

Context-Aware, AI-Powered ALT Text Generation for Lecture Visuals: A Multimodal Approach via Slide Decks and Speech Transcripts

Aybala Turkarslan

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Vainavi Thakkar

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Rodas Tekle

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Thomas Chen

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Jesse J. Martinez

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Jennifer Mankoff

Paul G. Allen School of Computer
Science and Engineering
University of Washington
Seattle Washington USA

Abstract

Blind or low vision (BLV) students face systemic digital barriers in STEM education. Among these barriers, a significant one is inaccessible lecture slides, particularly when ALT (alternative) text for visuals is missing or inadequate. ALT text can be onerous for experts to create for visual-heavy slide decks, and non-experts may not create informative ALT text. AI-powered generation techniques have promise to speed this process up, however AI text outputs can be inaccurate. To investigate this, we constructed a dataset of 30 lecture slide PDFs across 5 STEM domains. To generate ALT text for the lecture visuals, we developed a research-backed system prompt and internal AI-powered interface that uses lecture slide PDFs and slide-aligned transcripts. We ran an ablation study comparing image understanding of 795 lecture visuals under two research conditions: lecture slide PDF only and lecture slide PDF with additional transcript context. We scored the ALT text items across research-informed metrics such as Accuracy, Objectivity, Completeness, and Relevance. The AI-generated ALT text scored highly overall across all metrics in both research conditions, with minimal rates of hallucination. We conducted multiple comparisons, most of which were neutral, with some inconclusive effects after a Bonferroni correction for multiple comparisons. Our analysis did reject the null hypothesis in one case: for Illustrations, Completeness was +21.2% higher ($p < .001$) with transcript context. Across visual types, we also conducted a thematic analysis of comparative success and failure cases within AI-generated ALT text. Overall, our findings demonstrate the technical feasibility of AI-powered ALT text generation for STEM lecture visuals using multimodal context. These results motivate and lay the groundwork for future

BLV student user evaluations to further analyze how AI systems can support educational accessibility.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools; Accessibility technologies; Accessibility design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Education**.

Keywords

Multimodal large language models, generative AI, blind or low vision, alternative text, AI-generated alt text, context awareness, multimodal context, STEM lecture visuals, educational accessibility

ACM Reference Format:

Aybala Turkarslan, Vainavi Thakkar, Rodas Tekle, Thomas Chen, Jesse J. Martinez, and Jennifer Mankoff. 2026. Context-Aware, AI-Powered ALT Text Generation for Lecture Visuals: A Multimodal Approach via Slide Decks and Speech Transcripts. In *The 28th International ACM SIGACCESS Conference on Computers and Accessibility, October 25–28, 2026, Porto, Portugal*. ACM, New York, NY, USA, 35 pages. <https://doi.org/10.1145/3797867.3829048>

1 Introduction

Blind or low vision (BLV) students face systemic barriers in higher education, spanning areas such as campus navigation, inaccessible teaching materials, lack of instructor and institutional support, and peer stigma [4, 7, 13, 47]. As the prevalence of digital educational formats rises, the disparities faced by BLV higher education students are exacerbated in the form of inaccessible online lecture materials [3, 5, 24]. In particular, inaccessible lecture slide PDFs pose a serious visual accessibility barrier, as inaccessible digital documents can prevent screen readers from fully navigating content, announcing information, and explaining visual elements that lack descriptions [5, 8, 15]. For lecture materials that include many images and data visualizations, commonly found in STEM courses, these access barriers can greatly limit BLV students' access to course content with understandable visual information and opportunities for collaboration with peers [8, 34]. A result is that BLV students

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASSETS '26, Porto, Portugal

© 2026 ACM.

ACM ISBN 978-1-4503-XXXX-X/2025/10

<https://doi.org/10.1145/3797867.3829048>

are forced out of, or avoid pursuing, graphics-heavy academic and career pathways entirely [8].

It is the responsibility of universities, and instructors, to address this issue by creating proper image descriptions and supporting visual accessibility. However, numerous lawsuits demonstrate that such access is inconsistent at best [32, 33]. In the current digital landscape, such image description is primarily done through ALT (alternative) text. Writing high quality image ALT text can be slow, onerous, and difficult for non-experts who are often unsure of what details to include [3, 30]. Given this context, better quality automatically-generated image descriptions through the use of AI would be helpful to improve compliance. In addition, when BLV people have access to high quality AI image description generation, research shows that they benefit from and choose to use AI to ask questions and gather custom, person-specific information [45, 46]. Better tools that speed up this process could help ALT text content creators, provide a backstop for BLV students to access content, and potentially provide the ability to make “visual queries” similar to how non-BLV people explore diagrams [42], increasing equity in content access. However, image descriptions generated by large vision-language models (LVLMs) have been found to be factually inconsistent [14, 18, 22]. Recent studies have suggested that providing further context than just the image itself can improve accuracy [19, 25, 30]. However, none of these studies on accuracy or context within image descriptions focus on STEM lecture visuals.

A recent need-finding study with BLV participants showed that AI-assisted navigation and visual question-answering of lecture videos is a capability of high interest to this group [3]. We must understand how well AI functions in STEM settings, and what factors can influence this, to establish best practices and ensure that students using these technologies have equitable access to their lectures.

To address this gap, we develop a research-backed system prompt that can be used to generate ALT text for visuals in slide decks, using the entire lecture slide deck with slide-aligned, annotated lecture transcripts as context (described in Section 3). We evaluate the prompt as follows:

- We run an ablation study generating ALT text using AI for each visual on each lecture slide, using just the lecture slide PDF and the lecture slide PDF along with the annotated transcript. (Section 4)
- We evaluate a dataset of 795 images drawn from 30 lectures across 5 STEM topic fields to prove technical feasibility of AI-generated ALT text using a research-informed coding framework, grounded in prior BLV user study-centered literature. (Section 5)
- We explore the common reasons for AI-powered ALT text generation success and failure/error cases across lecture visual types, supplemented with comparative examples. We also provide a categorization of common types of visuals used in STEM lectures, showing that diagrams are by far most common (51%), despite being an underserved area of accessibility research [44]. (Section 5.1)

We conclude with a discussion (Section 6) of future work that is needed, including: BLV input into tool design, experiment replication across various AI models and lecture datasets, further refinement of the system prompt for ALT text generation, and coding schema co-development and evaluation studies with BLV students for AI-generated ALT text quality assessment.

2 Background on ALT Text and Generation Tools

ALT text is defined as “a brief text description of images and graphics” [41] used by assistive technologies such as screen readers. The World Wide Web Consortium (W3C) with its Web Content Accessibility Guidelines (WCAG) explains how to craft ALT text across a variety of non-decorative image types [12]. In this section, we review guidelines and research on ALT text quality, particularly for STEM content; and the use of AI to generate ALT text and the value of context in that process.

2.1 ALT Text Quality Guidelines

When determining guidelines for generating high quality ALT text, our research primarily focuses upon the perspectives of BLV participants across studies assessing the quality of image descriptions. Penuela et al. [18] found that BLV users of their MLLM (multimodal large language model)-enabled visual interpretation application were somewhat satisfied with its outputs and found it trustworthy, though text and graphics interpretation was error-prone with hallucinations (ie. false information generated by the model). The BLV participants of this study evaluated the application primarily upon *Accuracy*, *Satisfaction*, and *Trust*, and key design principles (of understanding user goals, maintaining a neutral tone, and correcting content quality) arose from the interviews [18]. Similarly for general image description, Garg et al. [14] found that their BLV study participants assessing the accuracy of VLM (vision language model)-generated photo image captions felt that the descriptions frequently lacked key details and contained inaccuracies. Surveyed BLV participants in this study across the board emphasized the need for *Precision/Specificity* and *Accuracy* in generated image descriptions, and highlighted the critical issue of how AI models are being “trained by non-disabled people, [and] show implicit bias toward disabled people” [14].

Similar perspectives are echoed amongst BLV participants in relevant studies assessing ALT text quality. Mack et al. [30] found that amongst their study’s BLV screen reader participants, the following ALT text traits were found to be necessary across all image types: *Accuracy* (ie. no incorrect information), *Completeness* (ie. describing all key aspects of an image within its context), and *Conciseness*. *Descriptiveness* and *Use of Natural Language* were also emphasized, though no participant prioritized naturalness over completeness [30]. Regardless of image type, many BLV participants in relevant studies emphasize the importance of the context an image lives within and its overall purpose in a file [30, 39]. Similarly, Mohanbabu and Pavel [19] found that BLV participants in their study assessing GPT-4V-generated context-aware web image descriptions appreciated the *Objectivity* and *Plausibility* of the outputs, though they noted difficulty in understanding unfamiliar context-specific terms and uncovered occasional hallucinations. Both Mack et al.

[30] and Mohanbabu and Pavel [19] find that one-size-fits-all ALT text quality guidelines (for length, tone, and terminology especially) are suboptimal, as BLV user preferences can vary across image type, user needs, personal prioritization of assessment criteria, and participant goals.

For data visualizations (such as charts, graphs, plots, etc.) which can commonly be found in STEM content, Lundgard and Satyanarayan [29] introduced a four-level model for the natural language semantic content description of such visualizations. Of the four levels, blind description readers in their study found visualization content and Levels 2 (statistical concepts and relations) and 3 (perceptual and cognitive phenomena) as most useful, while semantic content and Levels 1 (elemental and encoded structural properties) and 4 (contextual and domain-specific insights) were perceived as less useful [29]. Sharif et al. [35] arrived at similar conclusions, as BLV screen reader user participants of their study assessing visualizations using VoxLens (a multi-modal system that allows for verbal and audible interaction with visualizations) looked for holistic visualization information, performed “drilled-down explorations” [35] by comparing data points, and primarily sought to understand visualization trends and summary statistics. For data visualization ALT text in particular, Jung et al. [23] interviewed BLV participants who reviewed ALT text for existing visualizations and found that participants wish to create a mental model of the images, and that structural components of a chart (such as axes, range, ticks, chart type, and underlying data points) and the data trends were critical for user understanding. The majority of BLV participants in their study preferred an ALT text length ranging from 2-8 sentences, preferred plain terms, had varied opinions on the inclusion of color, and preferred the chart type to be stated first along with a summary of the visualization [23]. Similarly, Mack et al. [30] found that half of their BLV study participants believed visualization trends or summaries of takeaways should be included in ALT text, and a participant shared that current systemic barriers to visual accessibility explain “why you don’t see a lot of persons with disability going into very highly technical profession because of a lot of graphical illustration...are very, very inaccessible still” [30].

Some relevant works use referenceless quantitative metrics to assess image description quality, such as ClipScore (which analyzes a description’s quality compared to the image itself). Kreiss et al. [25] makes it clear that these cannot encapsulate context-sensitivity (i.e. the broader context in which an image appears), potentially making such metrics insufficient for image description quality evaluation. Previous work by Mohanbabu and Pavel [19] had BLV participants assess ALT text based on *Quality*, *Imaginability*, *Relevance*, and *Plausibility*; and Kreiss et al. [25] had BLV participants evaluate image descriptions across *Overall Quality*, *Imaginability*, *Relevance and Irrelevance*, and *Image Fit* (to its web article context). For data visualizations/informational graphics in particular, Smits et al. [37] identified design principles for accessible genomics data visualization descriptions with a focus upon BLV users: *Completeness*, *Accuracy*, *Conciseness*, *Structure*, *Navigability*, *Modularity*, and *Inclusivity*. Of all of these aforementioned metrics, though most were not developed on STEM image data specifically, they represent an important starting place for the evaluation of ALT text quality.

2.2 AI-Generated ALT Text

In recent years, the use of automated methods to generate ALT text has risen. This usage has spanned across informative images [31, 43] as well as data visualizations [6, 35, 37, 48]. With the latest developments in MLLMs in particular, generative AI has also increasingly been used for image description generation. However, studies with BLV participants have highlighted concerns with output accuracy and hallucinations [18], model difficulties with identifying low quality images accurately [14], results of low user satisfaction and trust [17], and a lack of reliability [2]. Further, for data visualizations, image descriptions generated by LVLs have been found to be factually inconsistent [22].

AI-generated ALT text can also be perceived as incomplete if it does not incorporate an image’s broader context. Studies by Kreiss et al. [25] and Mack et al. [30] underscore that the context that an image is found in holds an indispensable role in determining the usefulness of its description. Particularly for data visualizations, relevant studies assessing their visual accessibility emphasize the importance of providing context-aware, holistic semantic descriptions of both the imagery/structure as well as its underlying data, trends, and comparisons [10, 23, 29].

Therefore, context-aware ALT text generation tools have been developed, such as AltAuthor for accessible webpage development (which provides GPT-4o with an image, its image type, webpage context, and Section 508 accessibility guidelines to generate ALT text) [38] and ALADDIN for accessible educational materials in Word and PowerPoint files (which provides Gemini 2.0 Flash with an image, its textual content, document summaries, slide layout, image type, existing metadata, etc. to generate ALT text) [5]. When evaluated, automated context-aware image description generation has shown encouraging results. Mohanbabu and Pavel [19] found that context-aware descriptions for web images (generated using GPT-4V by providing the image and relevant details such as the webpage’s text, titles, etc.) were preferred by BLV participants in their study over context-free generated descriptions; and Penuela et al. [16] found that context-aware descriptions of real-world user photos (generated using Gemini 2.5 Pro by providing the image and a paired commonly-associated question from BLV users) led to better user need anticipation and human labeler preference. The potential impact of context-aware ALT text generation for STEM lecture visuals in particular, however, remains uninvestigated by such literature.

2.2.1 System Prompts for AI-Generated ALT Text. In building system prompts for AI-powered ALT text generation, related works begin by instructing the AI model with its goals before providing relevant rules [20, 36], primarily use zero-shot prompting [20, 31, 36], and provide short, intentionally structured prompts to mitigate “lost-in-the-middle” effects within long contexts [28, 31]. In crafting such system prompts, Mähr and Twente [31] and Singh et al. [36] draw upon relevant prior research [19, 40] pointing towards the improvement of visual descriptions by providing further relevant context for an image to justify incorporating such metadata into their prompting. Further, Mähr and Twente choose to place such metadata earlier in their prompt than the image input, to attempt to “effectively steer the model toward shape- and context-based reasoning rather than shallow texture correlations” [31]. These works

provide a useful foundation for the development of system prompts for AI-powered ALT text generation, but do not currently comprise of a methodology specific to STEM lecture visuals.

2.3 Multimodal Educational Datasets

While the aforementioned works were mainly not focused upon STEM lecture content, multimodal and context-aware understandings of digital lecture content is a growing research area (e.g., [9, 27]). However, the focus of such literature is generally not directly upon increasing accessibility for BLV students. For example, Lee et al. [27] explores the idea of providing automatic lecture context via automatic captions and education content generation, from information including spoken text, slide text, and visual figures. This dataset and overarching system can help improve the overall quality of information that students can access. Additionally, Chen et al. [9] introduces M^3AV , a large-scale multimodal dataset designed to support research on AI systems that can both recognize and understand academic lecture content by integrating additional context of speech, slides, and related papers. The potential for such work to translate into accessibility contexts is strong. This informational combination could provide rich explanatory context to students [11].

2.4 Summary

Prior studies of AI-generated ALT text have raised questions about quality and consistency, but have not addressed ALT text quality in the context of STEM education. In addition, prior work has demonstrated that the kind of context available during STEM lectures might be useful in addressing quality concerns. However, AI's value for ALT text generation has not been tested in the realm of lecture slide PDFs, nor specifically with STEM lecture materials. In this educational domain, increased context would typically be the lecture slides' visuals, corresponding text, and video speech transcript [9, 27]. Particularly for speech transcripts taken from lecture videos, work has already been done to synchronize and align lecture transcripts with slides [21]. Thus, the increased multimodal context combination of lecture slides and slide-aligned transcripts for enhancing AI-generated ALT text is an underexplored research area that warrants further investigation. Given the importance of context-aware high-quality ALT text and the increasing use of AI for its time-efficient generation, this paper investigates whether providing MLLMs with lecture slide PDFs and slide-aligned speech transcripts can improve the quality of ALT text generated for STEM lecture visuals.

3 Automating ALT Text Generation using AI

Generating ALT text from lecture slide PDFs and speech transcripts—particularly for STEM educational content—is a complex multimodal task requiring visual element interpretation, textual extraction and alignment, and contextual grounding across multiple long, science-heavy materials that may not be readily available on the Internet. Therefore, a model with strong capabilities in multimodal understanding and reasoning, scientific knowledge, and long context performance is vital.

To meet these needs, our team chose Gemini 3.1 Pro because of its multimodal capabilities and long context performance—as it is

positioned as a model “designed for tasks where a simple answer isn't enough”¹. We developed a system prompt for Gemini 3.1 Pro that could be used in conjunction with a PDF containing a slide deck, or a PDF plus annotated transcript, the two cases we test in our study.

Our system prompt was developed based upon our in-depth analysis of related work on building ALT text generation systems (described in Section 2.2) and BLV user evaluations of ALT text quality and AI-generated image description outputs (described in Section 2.1). When performing the ALT text generation API call to the Gemini model, our system prompt is first provided, then the annotated transcript data (if present), and finally the lecture slide PDF. This intentional design choice was based on work by Mähr and Twente [31], who place metadata earlier in their system prompt for ALT text generation than the image input, to attempt to “effectively steer the model toward shape- and context-based reasoning rather than shallow texture correlations” [31].

Our system prompt is found in Appendix A, and begins by informing the model of its role and goals towards generating high quality ALT text for lecture slide visuals (based upon work in Section 2.2.1). Our prompt also included a definition of what a visual object is, such that the outputted ALT text for both research conditions would result in the same amount of generated ALT texts per slide (ie. to avoid cases of one condition output perceiving multiple visuals in a slide as separate whilst another perceives them as all the same composite graphic, resulting in extra ALT text outputs for one of the conditions). Our prompt then includes information about how to handle lecture transcript context if provided, along with a JSON schema for our desired output of ALT text for each visual on each slide in a lecture PDF.

Finally, we end with key rules for the model to prioritize during generation (based upon Section 2.1), with an emphasis placed upon *Accuracy*, *Relevance*, *Completeness*, and *Specificity* as key metrics highlighted in relevant BLV user-centered evaluations [14, 18, 19, 30]. Due to variable BLV study participant preferences on *Conciseness versus Descriptiveness* [30], we simply instructed the model to provide ALT text outputs of 2-8 sentences in accordance to the BLV user preferences found by Jung et al. [23]. Additionally, we did not provide guidance on *Tone* or *Use of Natural Language* due to variable BLV user preferences on these criteria [19, 30]. For data visualizations in particular, we also instructed the model primarily in accordance with the aforementioned interview findings by Jung et al. [23] to place emphasis upon concepts along the lines of *Statistical Concepts and Relations*, *Perceptual and Cognitive Phenomena*, *Trends*, and *Structural Components* that were utilized in relevant BLV user evaluations [23, 29, 30, 35].

Our prompt also gave guidance on the appropriate use of context (to avoid hallucinations) and the importance of describing the entirety of repeated visuals across slides (rather than just minor changes within the visual), and instructed the model to avoid filler words and descriptions of decorative images. Finally, our prompt also worked to minimize the potential for error by instructing the

¹From a Google article by The Gemini Team on Gemini 3.1 Pro: 77.1% score on ARC-AGI-2 assessing a model's ability to solve entirely new logic problems; 80.5% score on MMMU Pro assessing multimodal task performance; and 84.9% score on MRCR v2 (8-needle) with 128k token context window size (average) assessing long context performance.

model to not hallucinate or guess uncertain content, as hallucinations can be common in AI-generated image description outputs [14, 18].

4 Study for Assessing ALT Text Quality

Using the system prompt described above, we conducted a technical feasibility study in regards to the quality of the resulting ALT text, and the impact of slide-aligned, annotated lecture transcript context on the quality of AI-generated ALT text for STEM lecture visuals. We used the ALT text outputs from our system prompt to run an ablation study, comparing AI-generated ALT text quality for the entire lecture slide PDFs with annotated lecture transcripts, to lecture slide PDFs alone. We describe our experimental setup, sample selection process, data preparation, coding process, and analytical approach below.

It is important to note that our research team does not include BLV individuals. To limit the impact of this, we focused our analysis on content-based and context-sensitive metrics that could be evaluated consistently while still capturing key aspects of ALT text quality, and we grounded our coding schema and evaluation criteria in prior studies that center BLV user perspectives (described in Section 2.1 and Section 2.2). This evaluation thus establishes evidence of feasibility, but is inherently incomplete. Our study lays the groundwork for future research to define evaluatory metrics and assess ALT text quality collaboratively with BLV end users.

4.1 Experimental Setup

Our study design consisted of 2 conditions (with transcript versus without transcript) $\times$ 5 fields (we included Anatomy, Biology, Dental, Machine Learning and Psychology) $\times$ 5 codes (plus 2 for data visualizations only) $\times$ 795 visual items. Items were individual visual elements on a slide (since while we ran our study on 30 lectures total, each had a variable number of slides, and there could be multiple items per slide).

To simplify data collection, we developed an AI-powered Chrome extension interface that allowed us to generate ALT text for STEM lecture visuals using our research-backed system prompt (described in Section 3). The internal Chrome extension interface enabled our experimenters to easily and rapidly upload lecture slide PDFs and slide-aligned transcripts, invoke the Gemini 3.1 Pro model through an API call using our system prompt, and retrieve structured JSON outputs containing AI-generated ALT text for lecture visuals. Our API call to the Gemini model was performed with a lower temperature (0.3) to work towards decreasing the amount of randomness in model token selection. The Chrome extension interface was implemented using Vite, React, TypeScript, and CSS.

4.2 Sample Selection

Our experimental setup required inputs to the AI model of lecture slide PDFs and corresponding slide-aligned, annotated lecture transcripts. Therefore, for our input data sample, we decided to use a multimodal dataset of educational lecture content² that covers a diverse range of topics, published by Lee et al. [27]. This dataset contains 180+ hours of video and 9000+ slides, with 10 lecturers from various subjects. The dataset has links to the lecture videos,

corresponding spoken language transcripts from Google ASR with timestamps, and manually annotated slide transition/segmentation timeframes. Each lecture is available on YouTube, and for each lecture slide in a lecture, there is the following associated metadata: the OCR of the in-slide text from Tesseract, the slide’s segmented spoken language, and the automatically extracted mouse traces. The educational content within the dataset primarily focuses upon STEM lectures and covers the following topic fields: Anatomy, Biology, Dental, Machine Learning, Psychology, and Speaking.

We sampled lectures for sufficiency, with the goal of identifying 30 lectures from at least 5 different topic fields that covered a range of visual types including illustrations, data visualizations, explanatory images, and diagrams. To find these lectures, we manually reviewed the lecture videos of the original dataset, eliminating those that only had decorative visuals or headshots of the lecturer. When we found a set of lectures that qualified, we selected 6 visual-heavy examples from that lecture set, and then moved on to look for a different field/lecture set. We excluded any lectures that had incomplete transcripts or other missing data from our final selection of 30 lectures. As a result of this, our team ended up having to discard two Biology lectures from our initial selection of 30 lectures. We chose to include the entire lecture PDF in every case, because that is the most realistic case for lecture ALT text generation contexts, where the full slide deck is usually available.

Our final sample included: 6 lectures from Anatomy, 6 lectures from Biology, 6 lectures from Dental, 6 lectures from Machine Learning, and 6 lectures from Psychology. The Anatomy lectures were taught at Ivy Tech Community College, the Biology lectures were taught at Los Medanos College, the Dental lectures were taught through Mental Dental (which prepares students for standardized dental exams), the Machine Learning lectures were taught at Carnegie Mellon University, and the Psychology lectures were taught at Keiser University. The specific lecture videos that were chosen from the original dataset can be found in Appendix B.

4.3 Data Preparation

We automated the process of deriving data that could be uploaded to our internal Chrome extension interface, which required a lecture PDF (generated using the original dataset’s lecture YouTube links and timestamps for lecture slide changes) and a slide-aligned, annotated transcript (generated using the original dataset’s lecture transcripts and timestamps for lecture slide changes). We also modified our interface slightly to log a JSON containing ALT text generated for each visual on each slide of a lecture PDF to the console for data collection purposes.

The unit of analysis was a single visual element within a lecture slide, such as a self-contained diagram, chart, image, or composite graphic that conveys a unified concept or message. The outputted ALT text data was organized into a table, where each row contained the lecture ID, slide number, visual ID, visual type, and the generated ALT text. For each visual on a lecture slide, the data table contains two rows, corresponding to the following conditions of AI-generated ALT text: *No Transcript Context* (ALT text generated using the lecture slide PDF only) and *Yes Transcript Context* (ALT text generated using the lecture slide PDF and the annotated transcript). Our process is summarized in Figure 1.

²The LPM Dataset GitHub repository is distributed under CC BY-NC-SA 4.0.

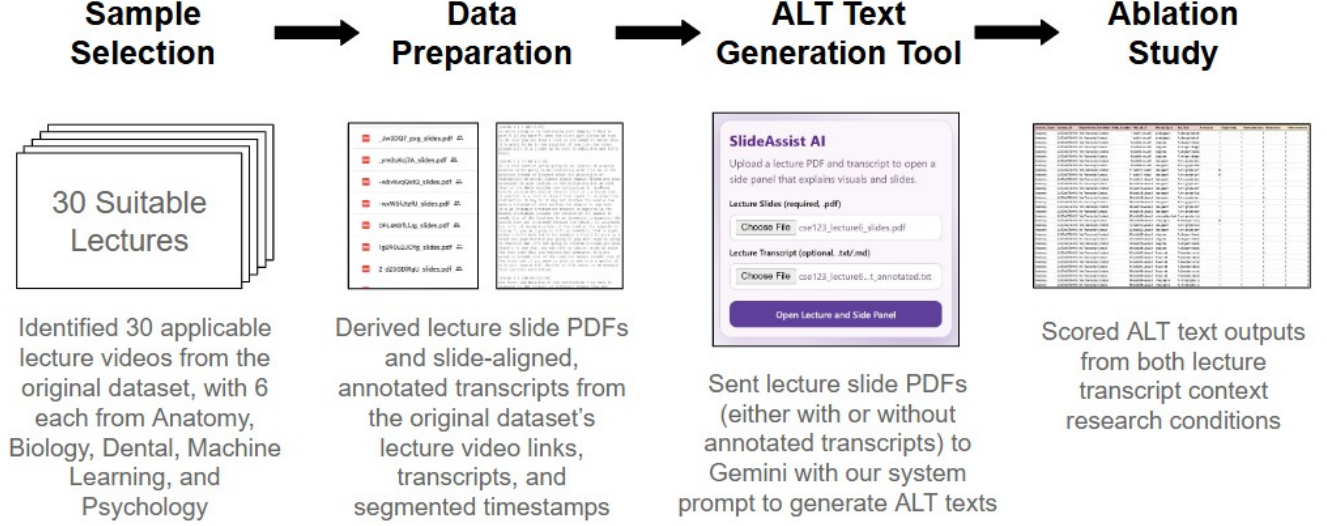


Figure 1: The process of deriving our data sample of 1590 ALT text items for the ablation study.

4.4 Data Coding and Inter-Coder Reliability

We developed a top-down codebook (defined in Appendix C) based on research-backed assessment metrics. When creating our codebook, our team drew from our analysis of research findings (in Section 2) from BLV users' assessments of ALT text and image description quality, including AI-generated outputs [14, 18, 19, 23, 29, 30, 35]. Our initial codebook included **Accuracy**, **Objectivity**, **Specificity**, **Completeness**, **Relevance**, **Presence of Hallucinations**, **Efficiency** (of length), and for visualizations only, **Structural Components** and **Data Relationships**. We decided to exclude **Satisfaction**, **Trust/Plausibility**, or **Imaginability** in our codebook criteria, as our coders are not BLV which would make these criteria particularly difficult to appropriately and accurately assess. We also excluded **Conciseness**, **Descriptiveness**, **Use of Natural Language**, **Tone**, and **Organization**, because preferences for these features vary significantly across BLV users. Specifically, Mack et al. [30] found that their BLV study participants prioritized content-based qualities of ALT text such as **Completeness** over **Use of Natural Language**, which informed our decision to focus further upon informational quality rather than stylistic delivery.

Our team independently coded the same Anatomy lecture and met to discuss disagreements and reshape our codebook and system prompt as needed. This led us to change our codes to all be binary instead of on a scale, to merge **Specificity** into **Completeness** and add more details into our codebook descriptions. We also removed **Efficiency** because the average sentence length of the ALT text generated was within the range of 2-8 sentences (average of 4.21 sentences with and 4.52 sentences without transcript context), which yielded no variability and aligned with related work on preferences [23]. We also updated the system prompt as a result of this discussion, to ensure that connected visuals on a slide are treated as a single unit, that all visuals on a slide are consistently described

without omissions, and that generated descriptions do not assume that the reader has prior context from previous slides.

Table 1: Inter-Rater Reliability (Final scores after both rounds. % (N) is percentage agreement (N=total items); α is Krippendorff's α . Note that the binary structure of our coding scheme and limited variability, due to high ratings across the board, influenced the interpretability of Krippendorff's α results for several codes. Thus, percentage agreement (i.e. the proportion of instances where coders all assigned the same code) was utilized as an important complementary metric that more directly reflects the high consistency observed.

Dimension	% (N)	α	Interp.
Accuracy	82 (22)	.46	Low-moderate
Objectivity	100 (22)	—	No variance
Completeness	82 (22)	.17	Very low
Relevance	100 (22)	—	No variance
Hallucinations	86 (22)	.23	Very low
Structural Components	86 (22)	.55	Moderate
Data Relationships	95 (22)	.33	Low

We then coded 3 lectures (10% of our dataset), and calculated inter-coder reliability using Krippendorff's α . After a second round of discussion and updates, we selected 11 example slide visuals to measure inter-coder reliability, which included 22 ALT text outputs, generated for a visualization-heavy Machine Learning lecture. The resulting Krippendorff's α ranged from .55 to .17, with very high exact percentage agreement (see Table 1). All codes achieved a percentage agreement (i.e. the proportion of instances where coders all assigned the same code) of $\geq 82\%$, indicating a high degree of consistency on the vast majority of scoring decisions. Given that

the binary structure of our coding scheme and limited variability (due to high ratings across the board) influenced the interpretability of Krippendorff’s α , this high percentage agreement stands as an important complementary metric that more directly reflects the high consistency observed among coders.

The full codebook, justification for each item, and scoring criteria can be found in Appendix C. For example, our process led to the following definition for **Accuracy**:

Whether all information in the ALT text is correct and supported by the visual (ie. “Is this an accurate/correct description of the visual?”).

Score 1 ONLY IF:

- All statements are supported by the visual (and transcript if applicable)
- No misleading or incorrect claims

Score 0 IF:

- Any incorrect fact or unsupported claim
- Any misrepresented relationship

In total, our team coded 1590 ALT text items across 30 lectures. To maintain coding consistency, for each lecture the assigned coder was tasked with coding both lecture transcript research condition outputs (ie. *No Transcript Context* and *Yes Transcript Context*) for each lecture slide visual. This helped to minimize coder difference as a cause of differing results between the two study conditions. Additionally, our team intentionally decided to split the ALT text items as evenly as possible between all 4 coders, and to split the lecture topics as evenly as possible as well. After coding all of the 30 lectures, our team also verified our coding consistency once more by selecting 6 example slides across 5 lectures (one for each domain of Anatomy, Biology, Dental, Machine Learning, and Psychology) to recalculate inter-coder reliability, which resulted in 60 ALT text outputs. The resulting Krippendorff’s α ranged from .78 to -.02, with very high exact percentage agreement (see the final region of Table 1).

4.5 Analytical Approach

To analyze the coded data, we scored conditions (*Yes Transcript Context* vs. *No Transcript Context*) using the proportion of ALT text that was rated as 1 on each score. We then calculated the proportion of visual elements that scored 1 on that metric, and the difference in proportions between conditions. These values are expressed as percentages in our results. We quantified uncertainty with 95% Confidence Intervals (bootstrap resampling, using 5,000 iterations).

Our analysis involves multiple comparisons across conditions and metrics. To mitigate this, we apply a Bonferroni correction within each group analyzed (i.e. when analyzing Data Visualizations across the 7 metrics that apply, we divide p values by 7). We also exclude cases with less than 10 total errors, or a 95% Confidence Interval (CI) that includes 0 (no difference).

5 Results

Overall, under our research-based evaluative framework, automated ALT text production achieved high scores across most metrics, as

shown in Figure 2. In either condition (with or without transcript context), most generated ALT text items received a score of 1 in all dimensions. The average score was over 95% for **Accuracy**, **Objectivity**, **Relevance** and **Hallucinations** in both conditions. This corresponds to a very low total number of errors. Errors [with transcript ($n=795$) and without transcript ($n=795$)] were [27, 22] for **Accuracy**, [4, 6] for **Objectivity**, and [11, 18] for **Relevance**. **Hallucinations** were very rare [7, 9].

In the case of **Completeness**, **Structural Components**, and **Data Relationships**, scores were slightly worse, but still quite good. Completeness showed high variability and was also the only metric for which a transcript made a significant difference, as discussed in Section 5.1. The number of errors were small for **Structural Components** [9, 5] and **Data Relationships** [4, 2] but the number of test cases was much smaller ($n=44$) and the CI much larger, making these differences less interpretable.

Table 2: Overall differences in ALT text quality with and without transcript context, across quality metrics. n = number of cases coded. T and NT refer to the transcript and no transcript context conditions. A positive difference indicates that having a transcript resulted in a percentage score that was that amount higher; negative indicates the reverse. Differences are reported as percentage-point gap in the proportion of items rated correct.

Quality Metric	n	T=0	NT=0	Diff (%)	95% CI	p
Accuracy	795	27	22	-0.6	[-2.0, 0.8]	.37
Objectivity	795	4	6	0.3	[-0.4, 0.9]	.41
Completeness	795	100	96	-0.5	[-3.4, 2.5]	.74
Relevance	795	11	18	0.9	[0.0, 1.8]	.05
Hallucinations	795	7	9	0.3	[-0.6, 1.1]	.56
Structural Comp.	44	9	5	-9.1	[-20.5, 0.0]	.10
Data Relat.	44	4	2	-4.5	[-13.6, 4.5]	.32

As mentioned in Section 4.5, to compare the lecture transcript context conditions across all metrics between visual pairs (i.e. the same ALT text output for a visual across both research conditions), we used bootstrap resampling to estimate 95% confidence intervals for the condition averages and average differences. For this overall analysis, no differences were significant (ie. none below $p=.05$). This could be because even in the *No Transcript Context* condition, the full lecture slide PDF was still provided to the AI model for ALT text generation. Full details are shown in Table 2. However, when we looked specifically at the impact of different visual types and lecture topic areas, we did find significant differences as described next in Section 5.1.

5.1 Visual Type Variation

We conducted an exploratory analysis in relation to the type of lecture visual being described. We identified seven distinct visual categories: **Data Visualization** ($n=44$ (6%): bar chart, composite chart, heatmap, line graph, pie chart, scatter plot, graph), **Diagram** ($n=408$ (51%): diagram, composite diagram, flowchart), **Composite** ($n=94$ (12%): composite graphic), **Photograph** ($n=154$ (19%): photograph, composite photograph, micrograph, screenshot, video frame, image grid), **Table** ($n=19$ (2%): table, table with diagram,

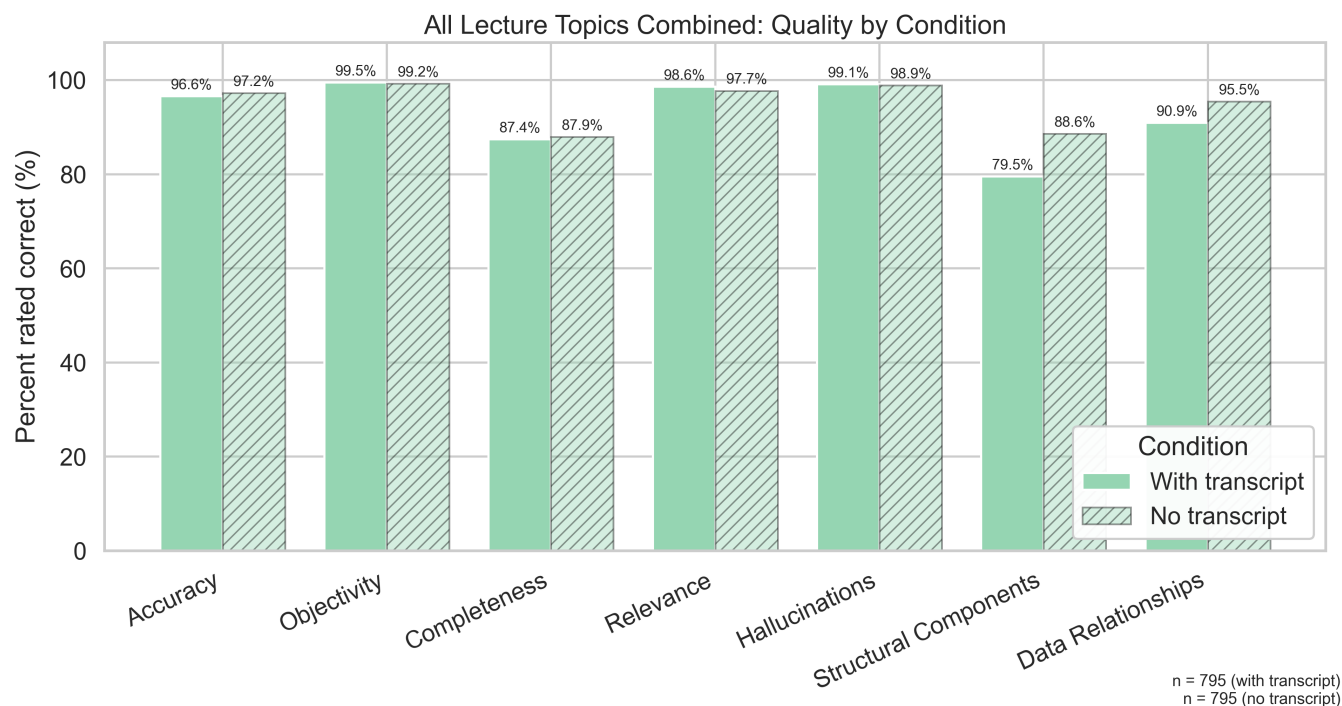


Figure 2: Barchart of overall performance showing very high average scores (above 95%) on almost all metrics, with high scores (80% or higher) on two metrics (Completeness and Structural Components).

table with photograph), **Illustration** (n=66 (8%): illustration, comic strip), and **Infographic** (n=10 (1%)). Only a few of these types have been emphasized in research on accessible ALT text. Given the small number of examples in some of these categories, future work should explore performance with an increased sample size.

In short, overall performance was high ($\geq 90\%$) on four of the five primary metrics, with and without transcript context. The least reliable metric was Completeness. Figure 3 shows how average performance varied on Completeness. Our team completed comparative success and failure case analyses detailed in Appendix D, where we define success as a case where ALT text generation resulted in a positive score on a metric when the transcript context was included, and a negative score when it was not (and failure as the reverse). We used thematic analysis to analyze potential success/failure causes, and found that when lectures were nonvisually accessible [26], ALT text quality improved with transcript context; while inaccessible lecturing practices could reduce ALT text quality given transcript context (see Figure 4). Full details for the assessed visual types can be found in Table 4, and the top five best and worst ALT text examples for each visual type can be found in Appendix E. We summarize our analysis of positive and negative transcript impacts here.

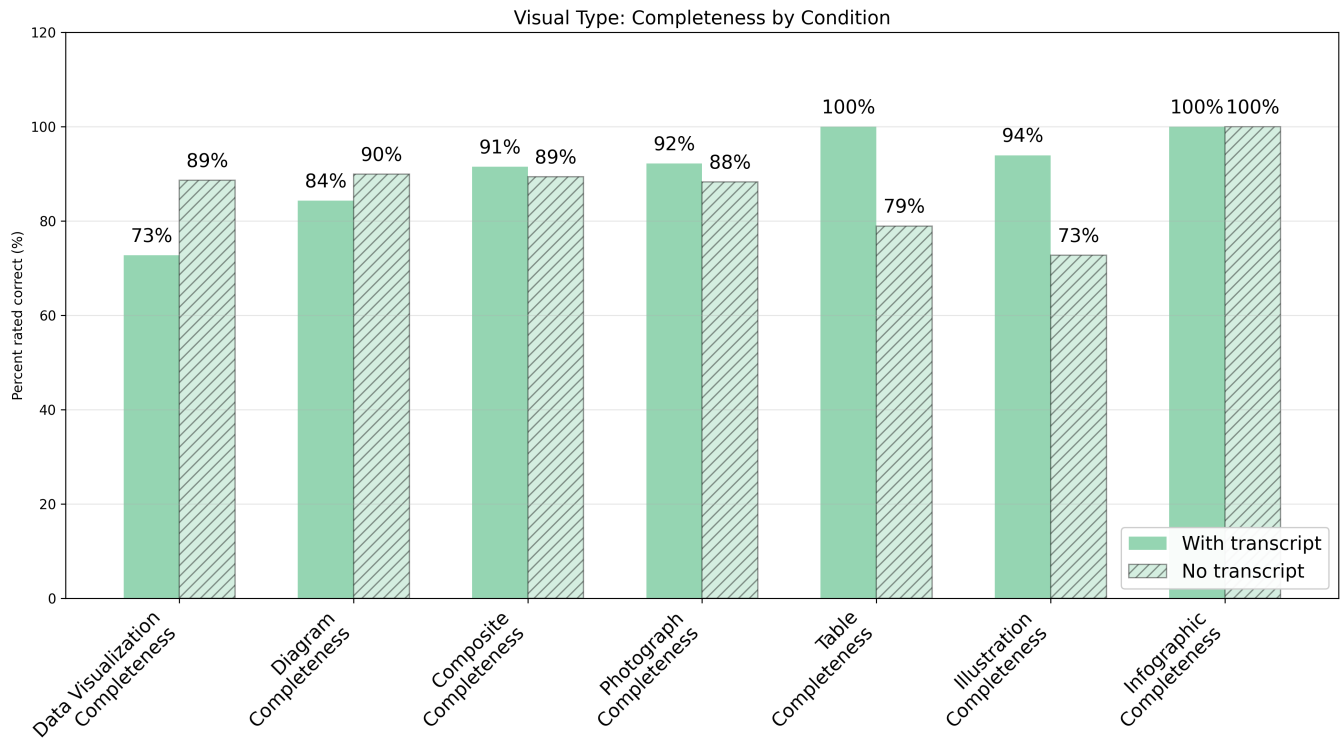


Figure 3: A barchart showing average performance on Completeness across visual image types.

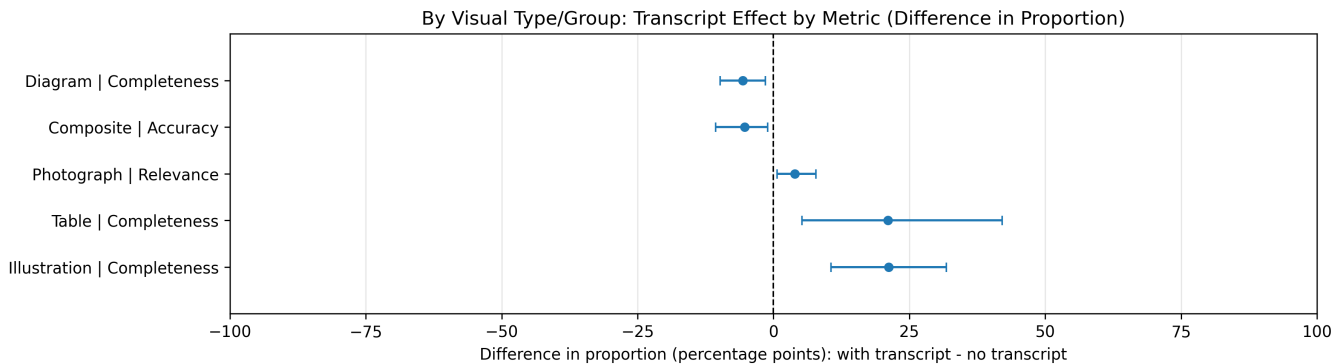


Figure 4: A forest plot depicting the effect of transcript context for multiple visual types. Bars show 95% CI intervals.

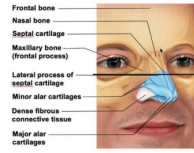
5.1.1 Positive Impact of Transcript. We found that transcripts which explained what a diagram represents or how it functions via additional semantic context, described more of an illustration’s existing content, or provided a richer description of a photograph while explaining its conceptual significance, improved ALT text output. We also found more comprehensive coverage of tabular information, and more comprehensive visual descriptions in transcripts that positively influenced ALT text. For example, in Figure 5. The largest positive change with transcript was in **Illustration | Completeness** (bottom line of Figure 4, 4 transcript errors vs 18 no

transcript errors, +21.2%, $p < .001$). Other positive changes were all borderline to varying degrees. A change in *Table | Completeness* (+21.1%) was caused by 4 no-transcript errors compared to 0 errors with transcript ($p = .04$). A small positive change with transcript was also observed in *Photograph | Relevance* (7 transcript errors vs 13 without transcript +3.9%, $p = .03$).

The Nose and Paranasal Sinuses (cont.)

External nose

- Areas include: **root** (area between eyebrows); **bridge**, *dorsum nasi* (anterior margin), and **apex** (tip of nose)
- **Nostrils (nares)**: bounded laterally by *alae*
- External nose is formed by:
 - Nasal and frontal bones superiorly—form bridge and root, respectively
 - Maxillary bones laterally
 - Plates of hyaline cartilage inferiorly



(a) Slide

An illustration showing the external skeletal framework of the nose. The right side of the image shows a person's face, while the left side shows the underlying bone and cartilage structure. Labels point to the frontal bone, nasal bone, septal cartilage, maxillary bone (frontal process), lateral process of septal cartilage, minor alar cartilages, dense fibrous connective tissue, and major alar cartilages. The root is the area between the eyebrows, the bridge is formed by the nasal bones, the *dorsum nasi* is the anterior margin, and the apex is the tip of the nose.

(b) ALT Text with transcript

An illustration of a human face showing the external skeletal framework of the nose. The right side of the nose is depicted with skin, while the left side reveals the underlying bones and cartilages. Labels point to the frontal bone, nasal bone, septal cartilage, maxillary bone (frontal process), lateral process of septal cartilage, minor alar cartilages, dense fibrous connective tissue, and major alar cartilages.

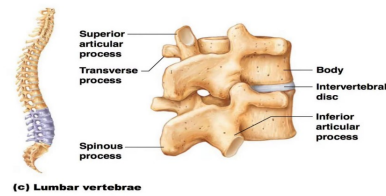
(c) ALT Text without transcript

Figure 5: For the positive impact of transcript context, an Anatomy lecture slide (slide 9, visual 1) with an **Illustration** that had more **Complete** ALT text with lecture transcript context through added spatial/structural information. Example text in the transcript that may have influenced this: *...the area between the eyebrow is called the route we have the bridge and then this is the the dorsal neighs I which is the also known as the anterior margin then finally we had the apex of the nose that's the tip of your nose ...*

© Original lecture provider; slide from LPM Dataset [27].

5.1.2 Negative Impact of Transcript. Transcripts that provided an overly high-level description of a diagram but omitted important descriptive details, or incorrectly identified or attributed items, negatively impacted ALT text. Other causes of worse ALT text included lack of information about visual emphasis, and incorrect mathematical or quantitative information. For example, one lecturer says “here is another schematic showing parts and pieces of non cutting and cutting instruments...so enough about that...” while in another case the lecturer says “let’s talk about blood for just a little bit,” which can steer the ALT text away from important information, causing reduced quality ALT text. For example, in Figure 6. Transcripts showed borderline negative changes on two metrics. A negative change with transcript was borderline significant for *Diagram | Completeness* (64 transcript errors vs 41 no transcript errors, -5.6% , $p = .01$) and for *Composite | Accuracy* (5 errors with transcript vs 0 without, -5.3% , $p=.02$).

Figure 7.21c Posterolateral views of articulated vertebrae.



(a) Slide

A diagram showing posterolateral views of articulated lumbar vertebrae. Labels point to the superior articular process, transverse process, spinous process, body, intervertebral disc, and inferior articular process.

(b) ALT Text with transcript

A posterolateral view of articulated lumbar vertebrae. The vertebrae feature massive bodies separated by intervertebral discs, short and thick spinous processes pointing straight backward, and transverse processes. Superior and inferior articular processes interlock adjacent vertebrae. A small inset shows the entire vertebral column with the lumbar region highlighted in purple.

(c) ALT Text without transcript

Figure 6: For the negative impact of transcript context, an Anatomy lecture slide (slide 50, visual 1) with a **Diagram** that had less **Complete** ALT text with lecture transcript context. The transcript primarily focused on the visual’s labels (instead of describing its appearance), which may have influenced this.

© Original lecture provider; slide from LPM Dataset [27].

5.2 Lecture Topic Variation

We investigated performance differences due to lecture topic across the 5 selected lecture domains: Anatomy, Biology, Dental, Machine Learning, and Psychology. There were no significant differences; and only a borderline difference across two metrics when transcript context was provided, as illustrated in Figure 7. *Anatomy | Completeness* increased from 20 with transcript to 33 errors without ($+9.1\%$, $p = .04$). *Biology | Accuracy* decreased from 9 errors with transcript 2 to errors without transcript (-3.8% , $p = .02$). Details for all of the assessed lecture topics can be found in Table 3.

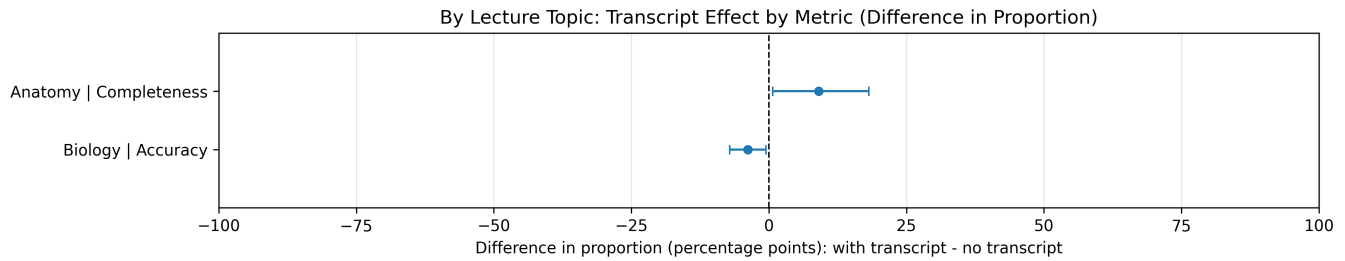


Figure 7: A forest plot depicting the effect of transcript context for multiple lecture topics.

6 Discussion

We have presented a technical feasibility study of context-aware, AI-powered ALT text generation for STEM lecture visuals. Our feasibility analysis examines whether AI generated ALT text is accurate, objective, complete, and relevant, with minimal hallucinations. Overall, the automatic ALT text generated using our AI system prompt scored highly in both lecture transcript context conditions across metrics, with hallucinations remaining rare. We also demonstrate that while most accessibility work focuses on data visualizations, the most common area of need is diagrams, followed by photographs, and then composite graphics. Our analysis was conducted by non-BLV individuals, and our metrics and AI system prompt were synthesized from published studies with BLV ALT text users. This work establishes the potential value of AI in this context. A critical next step is to evaluate accessibility outcomes directly with BLV users.

Our findings suggest that high-quality AI-generated ALT text is achievable for many STEM lecture visuals and that lecture transcript context can improve completeness for certain visual types. While transcript context improved ALT text completeness for certain visual types, in some cases additional context introduced errors, potentially due to shifting the model’s focus away from visual information or causing incorrect interpretations. These findings suggest that future AI-powered ALT text generation systems must carefully balance source visual material with added contextual information when generating descriptions for educational visuals. These findings highlight that transcript context can provide meaningful benefits for certain visual types, but may introduce tradeoffs for others.

6.1 Limitations

Our analysis is more comprehensive than any other study on AI-generated ALT text for STEM educational slide decks, containing a wide range of lecture visual types. Further, our findings highlight several opportunities and challenges for context-aware ALT text generation varying across educational visual types. However, our approach has several important limitations.

First, as already stated, our work did not include direct input from BLV individuals. It also did not include direct input from the authors of the slide decks in the public dataset by Lee et al. [27] that we evaluated. Thus, our metric ratings may be incorrect, from the perspective of our target audience. Further, we only measured aspects of ALT text that were not highly subjective, meaning that

important metrics found in the researched literature (ie. *Satisfaction*, *Trust/Plausibility*, *Imaginability*) were not tested by this study.

Further, the range of AI models and conditions that we tested is limited. Our evaluation uses only one AI model (Gemini 3.1 Pro) which limits the generalizability of our results across different multimodal AI systems. In addition, we provided significant context to the AI model (compared to typical single-image ALT text generation) in all conditions, since the entire lecture slide PDF was provided in every case (only the transcript presence varied). Future work could potentially compare these results to single-image ALT text generation. Additionally, further research could be done to include a wider variety of courses to capture the full diversity of visual materials used across educational disciplines.

6.2 Conclusions and Future Work

Overall, this work demonstrates the feasibility of AI-generated ALT text for STEM lecture visuals and identifies speech transcript context as a promising source of additional information for certain visual types. This technical feasibility study demonstrates the promise of the approach, however future research should involve BLV students throughout the entire ALT text generation and evaluation process. Collaborating directly with BLV students to co-develop an ALT text coding schema, and validate evaluation criteria is a critical next step. Further, a subsequent study should also investigate how AI-generated ALT text impacts lecture comprehension, navigation, satisfaction, and trust with BLV users. We should also study how BLV end users may respond to partially incomplete or mildly inaccurate AI-generated lecture visual descriptions, especially since context has been shown to impact error tolerance in other settings [1].

Additionally, our evaluation focused on one multimodal AI model and a selected set of STEM lecture materials. Future studies should examine additional AI models, disciplines, image complexities, datasets, and educational contexts to further understand the generalizability of these findings. Incorporating additional lecture context alongside speech transcripts (such as slide structure, figure annotations, etc.) could also potentially serve as a useful next step in improving context-aware, AI-powered ALT text generation for digital educational materials.

Further, the potential of variations in instructor narrative emphasis (in additional context such as lecture transcripts) to steer models away or towards generating high quality visual-focused ALT text also warrants future investigation. The balance between

model prioritization of source material (ie. the visual) versus additional context (ie. the slide containing the visual, the entire lecture slide deck, the lecture transcript, etc.) for AI-powered ALT text generation should be assessed further via tweaks to generation methodologies (ie. system prompts) to mitigate potential prompt injection effects from added context. Future research should also directly investigate the role of accessible lecturing practices on the quality of AI-powered ALT text generation when provided with lecture transcript context.

Acknowledgments

Generative AI was utilized to help write Python scripts to extract lecture slide PDFs and generate slide-aligned, annotated transcripts from the original dataset by Lee et al. [27]. It was used to generate scripts for calculating inter-coder reliability with Krippendorff's α and perfect agreement, and to help with the analysis script. Specifically, it was used to generate scripts that support data cleaning, perform statistical testing, and generate data visualizations. All of these scripts were verified manually and by writing tests to assess their accuracy. Generative AI was not used to write any of the text in this paper. Additionally, the contents of this article were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90REGE0026) funding the Center for Research and Education on Accessible Technology and Experiences (CREATE); and a research grant from GitHub Accessibility.

References

- [1] Ali Abdolrahmani, William Easley, Michele A. Williams, Stacy M. Branham, and Amy Hurst. 2017. Embracing : Examining How Context of Use Impacts Blind Individuals' Acceptance of Navigation Aid Errors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, CO, USA, May 06–11, 2017, Gloria Mark, Susan R. Fussell, Cliff Lampe, m. c. schraefel, Juan Pablo Hourcade, Caroline Appert, and Daniel Wigdor (Eds.). ACM, 4158–4169. doi:10.1145/3025453.3025528
- [2] Na Min An, Eunki Kim, Wan Ju Kang, Sangryul Kim, James Thorne, and Hyunjung Shim. 2025. How Blind and Low-Vision Individuals Prefer Large Vision-Language Model-Generated Scene Descriptions. arXiv:2502.14883 [cs.CV] <https://arxiv.org/abs/2502.14883>
- [3] Katharina Anderer, Karin Müller, Lukas Strobel, Matthias Wölfel, Jan Niehues, and Kathrin Gerling. 2025. Making Lecture Videos Accessible for Students who are Blind or have Low Vision through AI-Assisted Navigation and Visual Question Answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, Article 56, 15 pages. doi:10.1145/3663547.3746349
- [4] I Barcelona Bacalla, AL Gisultura, F Pacatang, R Romarate, and K Cabanilla Jr. 2024. Challenges and opportunities for students with visual impairment in a state university: a multi-case study. *International Journal of Multidisciplinary Research and Growth Evaluation* 5, 01 (2024), 1–10.
- [5] Disha Baglodi, Bella Martincic, Norah Sinclair, and Bruce N. Walker. 2025. Automated, Context-Aware Alt Text Generation for Educational Documents Using Large Language Models. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, Article 119, 4 pages. doi:10.1145/3663547.3759748
- [6] Abhijit Balaji, Thuvaaarakkesh Ramanathan, and Venkateshwarlu Sonathi. 2018. Chart-Text: A Fully Automated Chart Image Descriptor. arXiv:1812.10636 [cs.CV] <https://arxiv.org/abs/1812.10636>
- [7] Daniel Bishop and Daniel J.A. Rhind. 2011. Barriers and enablers for visually impaired students at a UK Higher Education Institution. *British Journal of Visual Impairment* 29, 3 (2011), 177–195. arXiv:<https://doi.org/10.1177/0264619611415329>
- [8] Matthew Butler, Leona Holloway, Kim Marriott, and Cagatay Goncu. 2016. Understanding the graphical challenges faced by vision-impaired students in Australian universities. *Higher Education Research and Development* 36 (05 2016), 1–14. doi:10.1080/07294360.2016.1177001
- [9] Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9041–9060.
- [10] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22)*. ACM, 1–12. doi:10.1145/3517428.3544796
- [11] Tu Anh Dinh, Philipp Nicolas Schumacher, and Jan Niehues. 2026. From Slides to Chatbots: Enhancing Large Language Models with University Course Materials. arXiv:2510.22272 [cs.CL] <https://arxiv.org/abs/2510.22272>
- [12] Eric Eggert, Shadi Abou-Zahra, and Brian Elton. 2022. Images Tutorial. <https://www.w3.org/WAI/tutorials/images/> World Wide Web Consortium (W3C), Updated 08 February 2022.
- [13] Tahsin Firat. 2021. Experiences of students with visual impairments in higher education: Barriers and facilitators. *British Journal of Special Education* 48, 3 (2021), 301–322.
- [14] Kapil Garg, Xinru Tang, Jimin Heo, Dwayne R. Morgan, Darren Gergle, Erik B. Sudderth, and Anne Marie Piper. 2026. "It's trained by non-disabled people": Evaluating How Image Quality Affects Product Captioning with VLMs. In *CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. <https://dl.acm.org/doi/10.1145/3772318.3791309>
- [15] Kawal Gill, Rekha Sharma, and Renu Gupta. 2017. Empowering visually impaired students through E-learning at higher education: problems and solutions. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)* 22, 8 (2017), 27–35.
- [16] Ricardo E. Gonzalez Penuela, Felipe Arias-Russi, and Victor Capriles. 2025. Guiding Multimodal Large Language Models with Blind and Low Vision Visual Questions for Proactive Visual Interpretations. In *Workshop on Vision Foundation Models and Generative AI for Accessibility: Challenges and Opportunities*. <https://openreview.net/forum?id=6vk6GpdMOo>
- [17] Ricardo E. Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 901, 21 pages. doi:10.1145/3613904.3642211
- [18] Ricardo E. Gonzalez Penuela, Crescentia Jung, Sharon Y. Lin, Ruiying Hu, and Shiri Azenkot. 2026. How Multimodal Large Language Models Support Access to Visual Information: A Diary Study With Blind and Low Vision People. In *CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. <https://dl.acm.org/doi/10.1145/3772318.3793266>
- [19] Ananya Gubbi Mohanbabu and Amy Pavel. 2024. Context-Aware Image Descriptions for Web Accessibility. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*. ACM, 1–17. doi:10.1145/3663548.3675658
- [20] Sabrina Haque and Christoph Csallner. 2025. Early Accessibility: Automating Alt-Text Generation for UI Icons During App Development. arXiv:2504.13069 [cs.SE] <https://arxiv.org/abs/2504.13069>
- [21] Alexander Holmberg. 2025. Generating Narrated Lecture Videos from Slides with Synchronized Highlights. arXiv:2505.02966 [cs.CV] <https://arxiv.org/abs/2505.02966>
- [22] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do LVLMS Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 730–749. doi:10.18653/v1/2024.findings-acl.41
- [23] Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2021. Communicating visualizations without visuals: Investigation of visualization alternative text for people with visual impairments. *IEEE transactions on visualization and computer graphics* 28, 1 (2021), 1095–1105.
- [24] Mohamed Koutheir Khribi. 2022. Toward accessible online learning for visually impaired and blind students. *Nafath* 6, 19 (2022).
- [25] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 conference on empirical methods in natural language processing*. 4685–4697.
- [26] Richard E. Ladner and Kyle Rector. 2017. Making your presentation accessible. *Interactions* 24, 4 (2017), 56–59. doi:10.1145/3085564
- [27] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2022. Multimodal Lecture Presentations Dataset: Understanding Multimodality in Educational Slides. arXiv:2208.08080 [cs.AI] <https://arxiv.org/abs/2208.08080>
- [28] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models

- use long contexts. *Transactions of the association for computational linguistics* 12 (2024), 157–173.
- [29] Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1073–1083. doi:10.1109/TVCG.2021.3114770
- [30] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) (ASSETS '21). Association for Computing Machinery, New York, NY, USA, Article 23, 14 pages. doi:10.1145/3441852.3471207
- [31] Moritz Mähr and Moritz Twente. 2025. Seeing History Unseen: Evaluating Vision-Language Models for WCAG-Compliant Alt-Text in Digital Heritage Collections. In *Computational Humanities Research 2025*, Taylor Arnold, Margherita Fantoli, and Ruben Ros (Eds.). Anthology of Computers and the Humanities, 1147–1167. doi:10.63744/njQVYcLndSPE
- [32] National Federation of the Blind. 2010. Penn State Discriminates Against Blind Students and Faculty. <https://nfb.org/about-us/press-room/penn-state-discriminates-against-blind-students-and-faculty>
- [33] National Federation of the Blind. 2012. National Federation of the Blind and Blind Student File Suit against Maricopa Community College District. <https://nfb.org/about-us/press-room/national-federation-blind-and-blind-student-file-suit-against-maricopa>
- [34] Ather Sharif, Sanjana Shivani Chintalapati, Jacob O. Wobbrock, and Katharina Reinecke. 2021. Understanding Screen-Reader Users' Experiences with Online Data Visualizations. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) (ASSETS '21). Association for Computing Machinery, New York, NY, USA, Article 14, 16 pages. doi:10.1145/3441852.3471202
- [35] Ather Sharif, Andrew M. Zhang, Katharina Reinecke, and Jacob O. Wobbrock. 2023. Understanding and Improving Drilled-Down Information Extraction from Online Data Visualizations for Screen-Reader Users. In *Proceedings of the 20th International Web for All Conference* (Austin, TX, USA) (W4A '23). Association for Computing Machinery, New York, NY, USA, 18–31. doi:10.1145/3587281.3587284
- [36] Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. 2024. FigurA11y: AI Assistance for Writing Scientific Alt Text. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) (IUI '24). Association for Computing Machinery, New York, NY, USA, 886–906. doi:10.1145/3640543.3645212
- [37] Thomas C Smits, Sehi L'Yi, Andrew P Mar, and Nils Gehlenborg. 2024. Alt-Gosling: automatic generation of text descriptions for accessible genomics data visualization. *Bioinformatics* 40, 12 (2024), btae670.
- [38] Hyungwoo Song, Minjeong Shin, Yeonjoon Kim, Kyochul Jang, Jaehoon Choi, Hyunggu Jung, and Bongwon Suh. 2025. AltAuthor: A Context-Aware Alt Text Authoring Tool with Image Classification and LMM-Powered Accessibility Compliance. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25 Companion)*. Association for Computing Machinery, New York, NY, USA, 124–128. doi:10.1145/3708557.3716366
- [39] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376404
- [40] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. arXiv:2307.05356 <https://arxiv.org/abs/2307.05356>
- [41] Harvard University. [n. d.]. Write helpful Alt Text to describe images. <https://accessibility.huit.harvard.edu/describe-content-images>
- [42] Colin Ware. 2005. *Visual Queries: The Foundation of Visual Thinking*. Springer Berlin Heidelberg, Berlin, Heidelberg, 27–35. doi:10.1007/11510154_2
- [43] Tracy Wei, Tiziano Piccardi, Miriam Redi, and Michael Bernstein. 2025. Bridging the Accessibility Gap: Generative AI for Alt Text on Wikipedia. In *Wiki Workshop (12th edition)*.
- [44] Brianna L. Wimer, Laura South, Keke Wu, Danielle Albers Szafir, Michelle A. Borkin, and Ronald A. Metoyer. 2024. Beyond Vision Impairments: Redefining the Scope of Accessible Data Representations. *IEEE Trans. Vis. Comput. Graph.* 30, 12 (2024), 7619–7636. doi:10.1109/TVCG.2024.3356566
- [45] Jingyi Xie, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M. Carroll. 2025. Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, Yokohama Japan, 26 April 2025– 1 May 2025*, Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas (Eds.). ACM, 62:1–62:17. doi:10.1145/3706598.3714210
- [46] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. 2024. Emerging Practices for Large Multimodal Model (LMM) Assistance for People with Visual Impairments: Implications for Design. arXiv:2407.08882 [cs.HC] <https://arxiv.org/abs/2407.08882>
- [47] Arun Kumar Yadav, A Yadav, and G Tripathi. 2024. Challenges faced by visually impaired students in inclusive higher education. *Educating for Societal Transitions* 27 (2024).
- [48] Chuqiao Yan, Hans-Peter Hutter, Felix M. Schmitt-Koopmann, and Alireza Darvishy. 2025. Chart Accessibility: A Review of Current Alt Text Generation. *IEEE Access* 13 (2025), 94040–94056. doi:10.1109/ACCESS.2025.3571626

A System Prompt

You are an accessibility assistant that generates high-quality alternative text (ALT text) for lecture slide visuals.

For each slide, detect meaningful visuals and generate ALT text for each one. If transcript context is provided, use it only to clarify visible content. If transcript context conflicts with the slide, trust the slide.

Return JSON only (no markdown, no extra text).

Schema: {

"slides": [

{

"slideNumber": number,

"visuals": [

{

"id": string

// (in the form of slide#-visual#),

"type": string

// (such as "diagram", "bar chart", "line graph", "scatter

// plot", "illustration", "photograph", "infographic",

// "flowchart", etc.)

"altText": string

}

]

}

]

}

Rules:

- slideNumber starts at 1 and increments by 1.
- If a slide has no visuals, return ``visuals``: [].
- One visual object per distinct visual. A distinct visual is a self-contained diagram, chart, image, or composite graphic that conveys a unified concept or message.
- If multiple visual elements on a slide are clearly connected or part of a single diagram, chart, or composite visual, treat them as ONE visual and describe all parts together in a single altText.
- altText must be accurate, relevant, complete, specific, and focused on instructional meaning (i.e. key elements, labels, relationships, directional indicators, and trends).
- altText should be within a range of 2-8 sentences as needed.
- When arrows, lines, or labels are present, explicitly describe what each one points to, where it is positioned relative to elements, and what relationship or flow it indicates (i.e. direction, sequence, or connection).
- If the visual is a data visualization or informational graphic (ex. chart, graph), begin altText by stating the chart type and a brief summary of what it shows.

Describe the axes and structural components, key data trends and relationships, and reference data labels or categories, highlighting any important comparisons or patterns that support understanding.

- You may weave brief relevant semantic context from prior slides and/or transcript into the literal visual description when it improves clarity.
- If transcript is present, use it when it improves precision or semantic meaning of visible elements (i.e. disambiguating acronyms, variable meaning, chart intent, or why a visual step matters).
- *Keep context tightly anchored to visible elements; do not add standalone context that is not grounded in what is on screen.*
- If a similar or repeated visual appears across multiple slides (even if slightly modified), treat each instance as independent. The altText must fully describe the visual as it appears on the current slide without assuming prior context or that the user has seen earlier slides. You may briefly note that the visual builds on or repeats a prior slide if helpful, but the description must remain fully understandable on its own.
- *Avoid filler/openers like "This image shows" or "This slide shows".*
- Write altText as direct description only (no bullets, no labels like ``ALT text:'', no markdown).
- Do not generate ALT text for a small video or image of the instructor unless it conveys instructional meaning (i.e. demonstrating an action, pointing to content, or containing relevant visual information). Otherwise, treat it as decorative.
- If purely decorative, set altText to `''`.
- If a value or label is not clearly visible, do not guess it.
- *If uncertain, describe only the clearly visible high-confidence elements.*
- Do not hallucinate details.

B Lectures Chosen

- **Anatomy**
 - Anatomy & Physiology Chapter 3 Part C Lecture
 - Anatomy and Physiology Chapter 7 Part B Lecture: The Skeleton
 - Anatomy and Physiology Chapter 7 Part C Lecture: The Skeleton
 - Anatomy & Physiology Chapter 9 Part B Lecture: Muscles & Muscle Tissue
 - Anatomy & Physiology Chapter 11 Part A: Nervous System & Nervous Tissue Lecture
 - Anatomy and Physiology Chapter 22 Part A Lecture: Respiratory System
- **Biology**
 - Human Biology Lecture: Ch 2 - Chemistry of Life
 - Human Biology Lecture: Ch 3 - Cell Structure and Function
 - Human Biology Lecture: Ch 6 - Reproduction at the Cellular Level Part 1
 - Human Biology Lecture: Ch 16 - The Digestive System

- Human Biology: Ch 16.3 - Cardiovascular System Part 1
- Human Anatomy Lecture: Ch 19 - The Heart Pt 1

- Dental

- Operative Dentistry | Composite Resin & Glass Ionomer | INBDE, ADAT
- Operative Dentistry | Dental Caries | INBDE, ADAT
- Prosthodontics | Clasp Design & Selection | INBDE, ADAT
- Head & Neck Anatomy | Cranial Nerves | INBDE
- Endodontics | Root Canal Treatment | INBDE, ADAT
- Basic Dental Instruments

- **Machine Learning**

- Lecture 2.2: Basic Concepts - Optimization (Multimodal Machine Learning, Carnegie Mellon University)
- Lecture 4.1: Multimodal Representations (Multimodal Machine Learning, Carnegie Mellon University)
- Lecture 4.2: Coordinated Representations (Multimodal Machine Learning, Carnegie Mellon University)
- Lecture 5.1: Multimodal Alignment (Multimodal Machine Learning, Carnegie Mellon University)
- Lecture 7.2: Generative Models (Multimodal Machine Learning, Carnegie Mellon University)
- Lecture 9.1: Reinforcement Learning (Multimodal Machine Learning, Carnegie Mellon University)

- **Psychology**

- PSY101 - States of Consciousness
- PSY101 - Personality
- PSY101 - Social Thinking
- Developmental Psychology - Middle Adulthood - CH11
- Developmental Psychology - Work, Leisure, & Retirement - CH12
- Developmental Psychology - Middle Adulthood - Biopsychosocial - CH13

C Final Codebook

The final codebook was as follows. We include the metric’s label, justification for including it, and the codebook definition.

Accuracy

Accuracy was frequently emphasized and evaluated in relevant assessment studies involving BLV participants [14, 18, 30]. BLV users consider ALT text correctness to be non-negotiable [30]. Inaccurate descriptions can be misleading and reduce trust in image descriptions, which is why measuring accuracy is integral to assessing whether generated ALT text reliably represents visual content. The codebook definition was as follows:

Whether all information in the ALT text is correct and supported by the visual (ie. “Is this an accurate/correct description of the visual?”).

Score 1 ONLY IF:

- All statements are supported by the visual (and transcript if applicable)
- No misleading or incorrect claims

Score 0 IF:

- Any incorrect fact or unsupported claim
- Any misrepresented relationship

Objectivity

Objectivity is an appreciated image description evaluation metric [19], that measures if the ALT text avoids subjective interpretation or unsupported assumptions. Neutral and factual descriptions are important as they can reduce bias that may misrepresent a visual and diminish description credibility. This metric ensures that descriptions are assessed on whether they remain grounded in the factual reality of a visual.

Definition: Whether the ALT text avoids interpretation not supported by the visual (ie. “Is this description void of assumptions?”).

Score 1 ONLY IF:

- All content is grounded in observable or stated information
- No speculation or interpretation beyond the source

Score 0 IF:

- Includes any assumptions or inferred meaning not shown
- Includes any subjective interpretation

Completeness

Completeness is an emphasized evaluation metric [30], that evaluates whether the ALT text includes all necessary information to understand a visual. Missing critical details can significantly reduce the usefulness of ALT text descriptions, especially when an image is complex. Therefore, we use this metric to help us assess whether descriptions provide a sufficient range of important information to both understand and mentally visualize an image.

Whether the ALT text includes the core information needed to understand and visualize the visual (ie. “Could I visualize the entire visual decently from this description?”).

Score 1 ONLY IF:

- A reader could understand and visualize the visual from the description alone, because it includes the main message and key details
- Includes important specific elements (ex. names, values, categories, etc.)
 - Does not need to mention colors unless they are vital to understanding the image
 - For very complex visuals (ex. with lots of labels, etc.), do not expect the description to necessarily capture every single thing, as long the vital elements are covered

Score 0 IF:

- A reader could NOT understand and visualize the visual from the description alone
 - Missing the main message
 - Missing key details
 - Too vague or generic

Relevance

Relevance is a driving evaluation metric for determining ALT text quality and usefulness [19, 25]. This metric is different from *Completeness*, since *Relevance* evaluates whether included information is connected to the instructional/lecture context, as opposed to *Completeness* which focuses on whether all key elements are covered to understand and mentally visualize an image. This metric also helps evaluate the presence of information that is important to the instructional context of the slide, which is key in lecture settings where visuals often serve a specific purpose.

Definition: Whether the ALT text explains the visual in a way that relates it to the lecture slide and the instructional context (ie. “Does this description explain how the visual relates to the slide?”).

Score 1 ONLY IF:

- The description explains the visual in relation to the lecture slide’s content
- Reflects the purpose of the visual within the slide
- Specifically, this can be met by there being at least 1-2 words from the slide text content reflected in the description

Score 0 IF:

- The description describes the visual without linking it to the lecture slide’s content

Presence of Hallucinations

Presence of Hallucinations was also evaluated, as fabricated and unsupported information is a prevalent issue within AI-generated content and can lead to user dissatisfaction and distrust [18, 19].

This metric identifies when the model introduces information that is not present in the visual or transcript, allowing us to more precisely analyze the output’s accuracy in the context of AI-driven automatic description generation.

Whether the ALT text includes information that is completely not present in the slide visual or lecture transcript. A hallucination is any statement that introduces new content that does not exist in the visual or transcript. This does NOT include misinterpretations of elements that are actually present (this would fall under Accuracy).

Score 1 ONLY IF:

- The description contains NO hallucinated content
- All mentioned elements are present in the visual or transcript

Score 0 IF:

- The description includes ANY hallucinated content
- Mentions objects, labels, or details that do not exist
- Introduces new elements not shown in the visual

Structural Components (visualizations only)

Structural Components captures the key structural elements of the data layout, which are important when BLV people interpret data visualization descriptions [23].

Score 1 ONLY IF:

- Identifies ALL key structural elements:
 - chart type
 - axes
 - labels
 - legend (if present)

Score 0 IF:

- ANY major structural element missing

Data Relationships (visualizations only)

Data Relationships supports the understanding of relationships between variables and overall data patterns, something BLV people prioritize when interpreting data visualization descriptions [23, 29, 30, 35].

Score 1 ONLY IF:

- Describes the main pattern(s)/relationship(s)/trends, such as:
 - Increase or decrease
 - Comparisons between groups
 - Correlations or associations
 - Notable overall patterns

Score 0 IF:

- Fails to describe the main pattern(s)/relationship(s)/trends in the data

Table 3: Difference in ALT text quality showing content type | metric, with and without transcript context, by visual categories. Rows with a significant p value are highlighted. T and NT refer to the transcript and no transcript context conditions. Differences are reported as percentage-point gap in the proportion of items rated correct. Rows with 0 difference are elided, or less than 10 items, are removed. We eliminated rows for quality metrics when there were 1 or 0 errors total, or no difference in accuracy, across conditions.

Quality Metric	n	T=0	NT=0	Diff (%)	95% CI	p
Anatomy Accuracy	143	2	5	+2.1	[0.0, 4.9]	.08
Anatomy Objectivity	143	0	2	+1.4	[0.0, 3.5]	.16
Anatomy Completeness	143	20	33	+9.1	[0.7, 18.2]	.04
Anatomy Relevance	143	0	1	+0.7	[0.0, 2.1]	.32
Anatomy Hallucinations	143	0	2	+1.4	[0.0, 3.5]	.16
Biology Accuracy	182	9	2	-3.8	[-7.1, -0.5]	.02
Biology Objectivity	182	1	3	1.1	[0.0, 2.7]	.16
Biology Completeness	182	13	10	-1.6	[-6.6, 3.8]	.53
Biology Relevance	182	5	6	0.5	[-1.1, 2.7]	.57
Biology Hallucinations	182	3	1	-1.1	[-2.7, 0.0]	.16
Dental Accuracy	129	6	5	-0.8	[-4.7, 3.1]	.71
Dental Objectivity	129	2	0	-1.6	[-3.9, 0.0]	.16
Dental Completeness	129	16	13	-2.3	[-10.1, 5.4]	.55
Dental Relevance	129	0	1	+0.8	[0.0, 2.3]	.32
Dental Hallucinations	129	1	0	-0.8	[-2.3, 0.0]	.32
ML Accuracy	238	6	5	-0.4	[-2.9, 2.1]	.74
ML Completeness	238	47	37	-4.2	[-10.5, 2.1]	.19
ML Relevance	238	1	3	0.8	[-0.8, 2.5]	.32
ML Hallucinations	238	3	6	1.3	[-0.8, 3.4]	.26
ML Structural Components	29	5	3	-6.9	[-17.2, 0.0]	.16
ML Data Relationships	29	2	1	-3.4	[-10.3, 0.0]	.33
Psych Accuracy	103	4	5	1.0	[-1.9, 3.9]	.57
Psych Completeness	103	4	3	-1.0	[-4.9, 2.9]	.66
Psych Relevance	103	5	7	+1.9	[-1.9, 5.8]	.32

Table 4: Difference in ALT text quality showing image type | metric, with and without transcript context, by visual categories. Rows with a significant p value are highlighted. T and NT refer to the transcript and no transcript context conditions. Differences are reported as percentage-point gap in the proportion of items rated correct. Rows with 0 difference are elided. We eliminated rows for quality metrics when there were 1 or 0 errors total, or no difference in accuracy, across conditions. *p* values are shown without the *Bonferroni correction*, but bolded only when still significant after correction.

Quality Metric	n	T=0	NT=0	Diff (%)	95% CI	<i>p</i>
Data Visualization Accuracy	44	2	4	+4.5	[0.0, 11.4]	.16
Data Visualization Completeness	44	12	5	-15.9	[-31.8, 0.0]	.05
Data Visualization Hallucinations	44	0	1	+2.3	[0.0, 6.8]	.32
Data Visualization Structural Components	44	9	5	-9.1	[-20.5, 0.0]	.10
Data Visualization Data Relationships	44	4	2	-4.5	[-13.6, 4.5]	.32
Diagram Accuracy	408	14	12	-0.5	[-2.5, 1.5]	.62
Diagram Objectivity	408	1	4	+0.7	[-0.2, 2.0]	.18
Diagram Completeness	408	64	41	-5.6	[-9.8, -1.5]	.01*
Diagram Hallucinations	408	6	8	+0.5	[-1.0, 2.2]	.53
Composite Accuracy	94	5	0	-5.3	[-10.6, -1.1]	.02
Composite Completeness	94	8	10	2.1	[-5.3, 9.6]	.60
Composite Relevance	94	2	4	+2.1	[0.0, 5.3]	.16
Composite Hallucinations	94	1	0	-1.1	[-3.2, 0.0]	.32
Photograph Objectivity	154	2	1	-0.6	[-1.9, 0.0]	.32
Photograph Completeness	154	12	18	+3.9	[-2.6, 10.4]	.24
Photograph Relevance	154	7	13	+3.9	[0.6, 7.8]	.03
Table Completeness	19	0	4	+21.1	[5.3, 42.1]	.04
Illustration Accuracy	66	1	2	+1.5	[-3.0, 6.1]	.57
Illustration Completeness	66	4	18	+21.2	[10.6, 31.8]	<.001*
Illustration Relevance	66	2	1	-1.5	[-7.6, 3.0]	.57
Infographic Accuracy	10	1	0	-10.0	[-30.0, 0.0]	.34

D Comparative Success & Failure Case Analyses

For each significant visual type and metric combination, ALT text generated (with and without transcript context) was qualitatively compared alongside the corresponding slide transcript. Recurring themes and ALT text examples are shown, along with indications of which successes and/or failures seem to have been directly influenced by a slide’s corresponding transcript information.

Table | Completeness

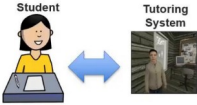
Our team completed a comparative success case analysis for the completeness of table ALT text, examining the 4 success cases (where the ALT text generated for a table visual was complete with the transcript but incomplete without the transcript context). No comparative failure cases (where the ALT text generated for a table visual was incomplete with the transcript but complete without the transcript context) were identified.

Success reasons given transcript context (n=4):

- **Theme 1: Explained what the visual represents via additional semantic context (n=1)**
 - Added Conceptual/Interpretive Information (n=1)
- **Theme 2: Described more of the visual’s existing content (n=3)**
 - Includes All Table Information (n=2)
 - Comprehensive Visual Description (n=1)

Example: Inferring Emotion from Interaction Logs

[Sabourin et al., 2011]



	Emotion Accuracy	Valence Accuracy
Baseline	22.4%	54.5%
Naive Bayes	18.1%	51.2%
Bayes Net	25.5%	66.8%
Dynamic BN	32.6%	72.6%

Language Technologies Institute
Carnegie Mellon University

(a) Slide

A table compares the performance of four models on emotion and valence accuracy for inferring student states from interaction logs. The models are Baseline, Naive Bayes, Bayes Net, and Dynamic BN. The Dynamic BN model achieves the highest accuracy in both categories, scoring 32.6% for Emotion Accuracy and 72.6% for Valence Accuracy. The Baseline model scores 22.4% and 54.5%, Naive Bayes scores 18.1% and 51.2%, and Bayes Net scores 25.5% and 66.8% respectively. Above the table is the illustration of the student interacting with the tutoring system.

(b) ALT Text with transcript

A table comparing the accuracy of different models for inferring emotion and valence. The models are Baseline, Naive Bayes, Bayes Net, and Dynamic BN. The Emotion Accuracy scores are 22.4%, 18.1%, 25.5%, and 32.6% respectively. The Valence Accuracy scores are 54.5%, 51.2%, 66.8%, and 72.6% respectively. The Dynamic BN model has the highest accuracy for both.

(c) ALT Text without transcript

Figure 8: For success Theme 2, a Machine Learning lecture slide (slide 17, visual 1) with a **Table** that had more **Complete** ALT text with lecture transcript context through “Comprehensive Visual Description”.

© Original lecture provider; slide from LPM Dataset [27].

Photograph | Relevance

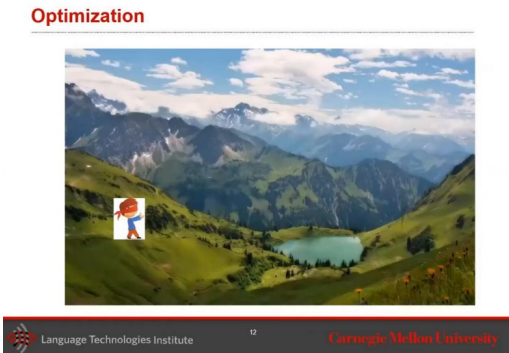
Our team completed comparative success and failure case analyses for the relevance of photograph ALT text, examining the 7 success cases (where the ALT text generated for a photograph visual was relevant with the transcript but not relevant without the transcript context) and the 1 failure case (where the ALT text generated for a photograph visual was not relevant with the transcript but relevant without the transcript context).

Success reasons given transcript context (n=7):

- **Theme 1: Explained how the visual relates to the surrounding course content (n=4)**
 - Explains Image-to-Slide Relationship (n=2)
 - Relates Image to Lecture Concept (n=2)
- **Theme 2: Provided a richer description of the visual while explaining its conceptual significance (n=3)**
 - Detailed Image Description with Conceptual Context (n=3; clearly based upon transcript information in 3/3 cases)

Failure reason given transcript context (n=1):

- **Theme 1: Failed to explain how the visual relates to the surrounding content (n=1)**
 - Missing Image-to-Slide Relationship (n=1)



(a) Slide

A photograph of a mountainous landscape in Switzerland with a small lake in the valley. A cartoon icon of a blindfolded person walking is superimposed on the left side of the image, illustrating the concept of optimization as exploring a landscape blindfolded to find the highest peak or lowest valley.

(b) ALT Text with transcript

A photograph of a mountainous landscape with a lake in the valley. A cartoon figure of a person with a red blindfold covering their eyes is superimposed on the left side of the image, standing on a grassy slope.

(c) ALT Text without transcript

Figure 9: For success Theme 2, a Machine Learning lecture slide (slide 6, visual 1) with a **Photograph** that had more **Relevant** ALT text with lecture transcript context through "Detailed Image Description with Conceptual Context". Example text in the transcript that may have influenced this: *...you want to find the highest peak in in Switzerland and so so you're walking around...in your optimization and you can see if you're if you're getting better or it's improving ...*

© Original lecture provider; slide from LPM Dataset [27].

Diagram | Completeness

Our team completed comparative success and failure case analyses for the completeness of diagram ALT text, examining the 27 success cases (where the ALT text generated for a diagram visual was complete with the transcript but incomplete without the transcript context) and the 50 failure cases (where the ALT text generated for a diagram visual was incomplete with the transcript but complete without the transcript context).

Success reasons given transcript context (n=27):

- **Theme 1: Explained what the visual represents or how it functions via additional semantic context (n=17)**
 - Added Conceptual/Interpretive Information (n=12; clearly based upon transcript information in 6/12 cases)
 - Added Technical/Scientific Terminology (n=5; clearly based upon transcript information in 5/5 cases)
- **Theme 2: Described more of the visual's existing content (n=9)**
 - Includes More Visual Details (n=4)
 - Includes All Table Information (n=2)
 - Includes Annotation/Label Information (n=2)
 - Includes Quantitative Information (n=1)

- **Undetermined (n=1)**

Failure reasons given transcript context (n=50):

- **Theme 1: High-level description that omitted important descriptive details (n=17)**
 - High-Level/Insufficient Detail Description (n=13; clearly based upon emphasis placed in transcript in 4/13 cases)
 - Label-Only Description (n=4; clearly based upon emphasis placed in transcript in 4/4 cases)
- **Theme 2: Failed to describe important information depicted in the visual (n=20)**
 - Missing Labels Information (n=9)
 - Missing Caption Information (n=4)
 - Missing Textual Content (n=4)
 - Missing Mathematical Information (n=3)
- **Theme 3: Failed to convey relationships or visual emphasis within the visual (n=10)**
 - Missing Visual Emphasis (n=6)
 - Missing Relational Information (n=4)
- **Undetermined (n=3)**

Structure and Function: Arteries

Structure

- Thick-walled
- Three layers



(a) Slide

A 3D cross-section diagram of an artery illustrating its thick-walled structure. The artery is composed of three distinct layers: an inner lining called the Tunica intima, a thick middle muscular layer called the Tunica media, and an outer protective layer called the Tunica externa. The central empty space where blood flows is the lumen.

(b) ALT Text with transcript

A 3D cross-section diagram of an artery showing its thick-walled, three-layer structure. The outermost layer is a tough protective sheath, the middle layer is thick smooth muscle, and the innermost layer is a smooth lining surrounding the hollow central lumen.

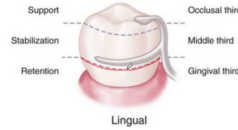
(c) ALT Text without transcript

Figure 10: For success Theme 1, a Biology lecture slide (slide 18, visual 1) with a **Diagram** that had more **Complete** ALT text with lecture transcript context through "Added Technical/Scientific Terminology". Example text in the transcript that may have influenced this: *...innermost layer is the Tunica intima ...this middle muscular layer which is the Tunica Media and then the outermost layer of the artery is the Tunica externa...*

© Original lecture provider; slide from LPM Dataset [27].

Reciprocal Clasp (Stabilizing Clasp)

- Originates from minor connector and rest
- Contacts tooth **above** HOC/survey line
- Braces abutment tooth so it is not torqued by retentive clasp



(a) Slide

An illustration of a tooth from a lingual view demonstrating the placement of a reciprocal or stabilizing clasp. The tooth is divided into occlusal, middle, and gingival thirds, with a solid red line indicating the height of contour. The reciprocal clasp originates near the occlusal surface and extends horizontally across the middle third of the tooth. Unlike a retentive clasp, the entire reciprocal clasp arm, including its tip, remains completely above the red height of contour line to brace the abutment tooth without engaging the undercut.

(b) ALT Text with transcript

A diagram of a tooth illustrates the position of a reciprocal, or stabilizing, clasp on the lingual surface. A dashed purple line marks the height of contour across the middle third of the tooth, and a solid red line marks the undercut area in the gingival third. Unlike the retentive clasp, the entire reciprocal clasp arm is positioned above the height of contour, resting entirely within the middle third stabilization zone. Labels on the left indicate the functional zones: 'Support' in the occlusal third, 'Stabilization' in the middle third, and 'Retention' in the gingival third.

(c) ALT Text without transcript

Figure 11: For failure Theme 2, a Dental lecture slide (slide 5, visual 1) with a **Diagram** that had less **Complete** ALT text with lecture transcript context through "Missing Labels Information". The transcript lacked any mention of the labels within the visual, which may have influenced this.

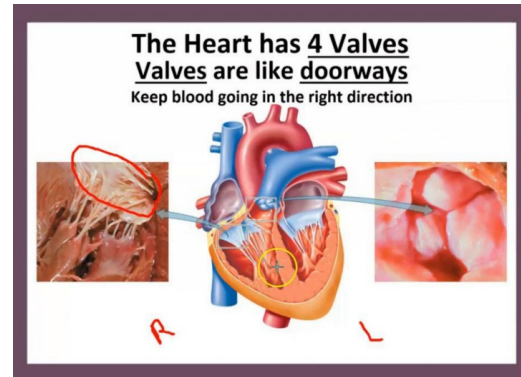
© Original lecture provider; slide from LPM Dataset [27].

Composite | Accuracy

Our team completed a comparative failure case analysis for the accuracy of composite ALT text, examining the 5 failure cases (where the ALT text generated for a composite visual was inaccurate with the transcript but accurate without the transcript context). No success cases were identified (i.e. there were no cases where the ALT text generated for a composite visual was accurate with the transcript but inaccurate without the transcript context).

Failure reasons given transcript context (n=5):

- **Theme 1: Incorrectly represented relationships between elements in the visual (n=2)**
 - Missing Relational Information (n=2)
- **Theme 2: Incorrectly identified or attributed items depicted in the visual (n=3)**
 - Incorrect Object Identification (n=1; error likely based on emphasis in transcript)
 - Incorrect Labels Information (n=1)
 - Incorrect Quantitative Information (n=1)



(a) Slide

A composite visual showing the four valves of the heart. In the center is a cross-section diagram of the heart with arrows indicating the forward flow of blood through the chambers and valves. A yellow circle highlights the right atrioventricular valve between the right atrium and right ventricle. To the left, an arrow points to a photograph of an open heart valve, with its string-like chordae tendineae circled in red. To the right, an arrow points to a photograph of a closed valve viewed from above, displaying its cusps tightly sealed together. Handwritten red letters 'R' and 'L' denote the right and left sides of the visual.

(b) ALT Text with transcript

A central cross-section diagram of the heart highlights its internal valves. An arrow points from the right side of the heart to an inset photograph on the left, labeled 'R', showing the string-like chordae tendineae attached to a valve. Another arrow points from the left side of the heart to an inset photograph on the right, labeled 'L', showing a close-up of a closed, three-flap valve resembling a doorway.

(c) ALT Text without transcript

Figure 12: For failure Theme 2, a Biology lecture slide (slide 10, visual 1) with a **Composite** that had less **Accurate** ALT text with lecture transcript context through "Incorrect Object Identification" where the location of the yellow circle is described incorrectly. Example text in the transcript that may have influenced this: *...and The ventricle this is called the atrioventricular valve okay we also have an atrial ventricular valve on the left...*

© Original lecture provider; slide from LPM Dataset [27].

Illustration | Completeness

Our team completed comparative success and failure case analyses for the completeness of illustration ALT text, examining the 15 success cases (where the ALT text generated for an illustration visual was complete with the transcript but incomplete without the transcript context) and the 1 failure case (where the ALT text generated for an illustration visual was incomplete with the transcript but complete without the transcript context).

Success reasons given transcript context (n=15):

- **Theme 1: Described spatial or structural relationships within the visual (n=7)**
 - Added Spatial/Structural Information (n=7; clearly based upon emphasis placed in transcript in 7/7 cases)
- **Theme 2: Described more of the visual's existing content (n=7)**
 - More Detailed Visual Description (n=4; clearly based upon emphasis placed in transcript in 3/4 cases)
 - Mentions Repeated Image (n=3)

- **Theme 3: Explained what the visual represents via additional semantic context** (n=1)
 - Added Conceptual/Interpretive Information (n=1)

Failure reasons given transcript context (n=1):

- **Theme 1: Failed to include textual information associated with the visual** (n=1)
 - Missing Textual Content (n=1)

E Top 5 Best & Worst ALT Text Examples (with transcript context)

For each visual type, our team paired the two versions of the same ALT text item (for the same lecture slide visual): one generated with transcript context and one without transcript context. For every pair, we compared the available evaluation metrics (ie. Accuracy, etc.), computed a with-transcript mean score and a no-transcript mean score across the metrics, and then computed the difference as ($yes_{mean} - no_{mean}$).

Within each visual category, “best” examples are marked 1+ (1st) to 5+ (5th), and were those with positive differences, ranked highest by with-transcript mean (then by larger positive difference and count of positively-scored metrics as tie-breakers), while “worst” examples are marked 1- (1st) to 5- (5th), and were those with negative differences, ranked lowest by with-transcript mean (then by larger negative difference and count of negatively-scored metrics as tie-breakers). The top five best/worst ALT text examples given transcript context for each visual type can be found here.

Data Visualization

Difficulty of Dentin Bonding

- **Composition**—enamel is about 90% mineral, dentin has much more organic matter and water
- **Structure**—enamel rods are parallel, collagen is a “bowl of spaghetti”
- **Depth**—fluid-filled dentinal tubules are much larger and more numerous near pulp, so less bond strength the deeper you are
- **Smear layer**—“sawdust” from cutting, decreases dentin permeability

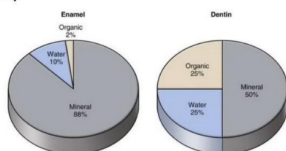
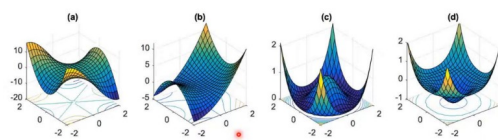


Figure 13: 1+: Dental (4:1)

© Original lecture provider; slide from LPM Dataset [27].

Bad Saddle Points



<https://arxiv.org/pdf/1602.05908.pdf>

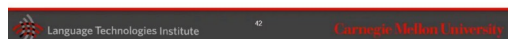


Figure 14: 2+: ML (33:1)

© Original lecture provider; slide from LPM Dataset [27].

Example

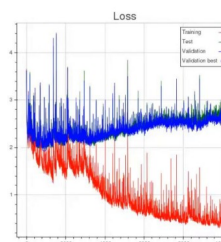


Figure 15: 3+: ML (35:1)

© Original lecture provider; slide from LPM Dataset [27].

Example

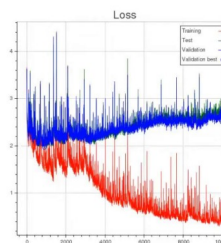


Figure 16: 4+: ML (37:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 5: ALT text for Data Visualization Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	Two pie charts comparing the composition of enamel and dentin. The left chart shows enamel is composed of 88% mineral, 10% water, and 2% organic material. The right chart shows dentin is composed of 50% mineral, 25% water, and 25% organic material. The charts illustrate that dentin has significantly less mineral and much more organic matter and water than enamel.
1+	N	Two pie charts comparing the composition of enamel and dentin by volume. The first pie chart shows enamel is composed of 88% Mineral, 10% Water, and 2% Organic material. The second pie chart shows dentin is composed of 50% Mineral, 25% Water, and 25% Organic material.
2+	Y	Four 3D surface plots labeled (a), (b), (c), and (d), illustrating different types of 'Bad Saddle Points'. Plot (a) shows a standard saddle point. Plot (b) shows a steeper saddle point. Plot (c) shows a complex surface with multiple peaks and valleys, including a saddle point. Plot (d) shows a surface with a flat region that transitions into a saddle point. These illustrate the complex topography that optimization algorithms must navigate in deep learning.
2+	N	Four 3D surface plots labeled (a) through (d) showing different types of saddle points. Plot (a) is a standard saddle point. Plot (b) is a steeper saddle point. Plots (c) and (d) show more complex surfaces with multiple peaks and valleys, illustrating 'bad' saddle points where optimization algorithms might get stuck. A citation at the bottom reads 'https://arxiv.org/pdf/1602.05908.pdf'
3+	Y	A line graph showing training and validation loss over 10,000 iterations. The x-axis is iterations and the y-axis is loss. The red line represents 'Training' loss, which steadily decreases towards zero, indicating the model is learning the training data well. The blue line represents 'Validation' loss, which initially decreases but then starts to increase after around 2000 iterations, indicating overfitting. The green line represents 'Test' loss, which follows a similar pattern to the validation loss.
3+	N	A line graph showing training and validation loss over 10,000 iterations. The x-axis is iterations and the y-axis is Loss. The red line represents training loss, which steadily decreases towards zero. The blue line represents validation loss, which initially decreases but then starts to increase after about 2000 iterations, indicating overfitting.
4+	Y	A line graph showing training and validation loss over 10,000 iterations. The x-axis is iterations and the y-axis is loss. The red line represents 'Training' loss, which steadily decreases towards zero, indicating the model is learning the training data well. The blue line represents 'Validation' loss, which initially decreases but then starts to increase after around 2000 iterations, indicating overfitting. The green line represents 'Test' loss, which follows a similar pattern to the validation loss.
4+	N	The same line graph from slide 35, showing training loss decreasing and validation loss increasing over time, indicating overfitting.
5+	Y	Stacked bar chart titled 'The Top 10 Elements Found in Your Body'. The chart uses a human silhouette to represent the proportions of elements. The 'Big 4' elements make up 96% of the body: Oxygen at 65%, Carbon at 18.5%, Hydrogen at 9.5%, and Nitrogen at 3%. The remaining 4% consists of other elements like Calcium, Phosphorus, Potassium, Sulfur, Sodium, and Chlorine, along with trace amounts of 15 other elements.
5+	N	A bar chart titled 'The Top 10 Elements Found in Your Body' showing the percentage of body composition. A human figure is filled with colors corresponding to elements: Oxygen (65%, red), Carbon (18.5%, dark gray), Hydrogen (9.5%, light gray), and Nitrogen (3%, blue). These four make up 96% of the body. A small section at the bottom represents 'Other (4%)', which includes Calcium, Phosphorus, Potassium, Sulfur, Sodium, Chlorine, and trace amounts of 15 other elements.

Table 6: ALT text for Data Visualization Worst Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1-	Y	A diagram showing an alternative matrix formulation of Dynamic Time Warping. On the left, the two unaligned signals (red and blue) are shown with vertical arrows indicating where points are replicated to align the sequences. On the right, this alignment is represented mathematically as the original signals (X and Y) multiplied by binary alignment matrices (Wx and Wy). The matrices contain 1s and 0s, where the 1s define the alignment mapping, allowing the objective to be formulated using the Frobenius norm.
1-	N	A diagram showing an alternative formulation of Dynamic Time Warping using alignment matrices. On the left, two line graphs show the original unaligned signals (red and blue) and their aligned versions with dashed lines connecting corresponding points. Red arrows point down to replicated points in the red signal, and blue arrows point up to replicated points in the blue signal. On the right, the alignment is represented by matrix multiplication. The original red signal is multiplied by a binary alignment matrix X, and the original blue signal is multiplied by a binary alignment matrix Y. The matrices X and Y contain 1s and 0s, indicating which points in the original signals correspond to the aligned sequence. The text notes 'Replication doesn't change the objective!' and presents the alternative objective function using the Frobenius norm.
2-	Y	A line graph demonstrates unfused or incomplete tetanus due to low stimulation frequency. The x-axis represents Time in milliseconds, and the y-axis represents Tension. Multiple stimuli are applied, indicated by upward red arrows. If another stimulus is applied before the muscle relaxes completely, more tension results. This wave or temporal summation causes the tension to increase with each successive stimulus, creating a wavy curve that rises progressively higher, showing partial relaxation between stimuli.
2-	N	A line graph illustrating unfused, or incomplete, tetanus due to low stimulation frequency. The y-axis represents tension, and the x-axis represents time in milliseconds from 0 to 300. Four stimuli are applied at regular intervals. After the first stimulus, the muscle contracts and begins to relax. Before it fully relaxes, the next stimulus is applied, causing the tension to rise higher than the previous peak. This pattern of partial relaxation followed by increased tension, known as wave summation, continues with each subsequent stimulus, resulting in a jagged, rising tension curve.
3-	Y	A line graph demonstrates fused or complete tetanus due to high stimulation frequency. The x-axis represents Time in milliseconds, and the y-axis represents Tension. A rapid succession of stimuli is applied, indicated by many closely spaced upward red arrows. At higher stimulus frequencies, there is no relaxation at all between stimuli. The tension rises smoothly to a maximum plateau, creating a sustained contraction known as fused tetanus, before eventually declining.
3-	N	A line graph showing fused, or complete, tetanus resulting from high stimulation frequency. The y-axis represents tension, and the x-axis represents time in milliseconds from 0 to 300. Numerous stimuli are applied in rapid succession. The muscle tension rises smoothly and continuously without any periods of relaxation between stimuli, reaching a sustained maximum tension plateau. Once the rapid stimuli cease, the tension drops back to zero.
4-	Y	Two line graphs illustrating Dynamic Time Warping (DTW). The top graph shows two unaligned temporal signals: a red line (X) and a blue line (Y), both with 8 data points plotted on a y-axis from 1 to 5. The bottom graph shows the same two signals, but with dashed black lines connecting points on the red line to corresponding points on the blue line. These connections represent the set of indices chosen to minimize the alignment difference between the two sequences.
4-	N	Two line graphs illustrating Dynamic Time Warping (DTW) for unimodal signals. The top graph shows two unaligned temporal signals: a red line (signal X) and a blue line (signal Y), both plotted against time indices 1 through 8. The red signal has peaks and valleys that are out of sync with the blue signal. The bottom graph shows the same two signals, but with dashed black lines connecting corresponding points between the red and blue signals. These connections represent the alignment found by DTW to minimize the difference between the two signals, showing how points in one signal are mapped to points in the other, sometimes mapping one point to multiple points to account for temporal shifts.
5-	Y	A line graph titled 'Stephan Curve (FA)' for fluorapatite. The axes are the same as previous Stephan curves, showing Plaque pH over Time. The curve drops from pH 7 to about 4.5 at 10 minutes and then recovers. However, the horizontal line for 'Critical pH' is lowered to 4.5. Because the curve never dips below this new, lower critical pH, the entire area under the curve is shaded green, indicating no demineralization occurs.
5-	N	A line graph titled 'Stephan Curve (FA)' showing Plaque pH over 60 minutes. The x-axis represents Time from 0 to 60 minutes, and the y-axis represents Plaque pH from 4.5 to 7. A horizontal black line is positioned at the very bottom of the y-axis at pH 4.5, labeled 'Critical pH'. The black curve drops sharply from pH 7 to just above 4.5 at 10 minutes, then gradually rises back to 6.5 by 60 minutes. Because the entire curve remains above the lowered critical pH line of 4.5, the entire area under the curve is shaded green, indicating no demineralization occurs.



Figure 17: 5+: Bio (4:2). Note that a human individual's face on the slide has been manually blocked out by our team for privacy.

© Original lecture provider; slide from LPM Dataset [27].

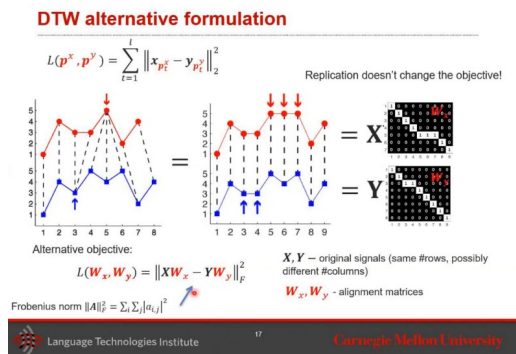


Figure 18: 1-: ML (11:1)

© Original lecture provider; slide from LPM Dataset [27].

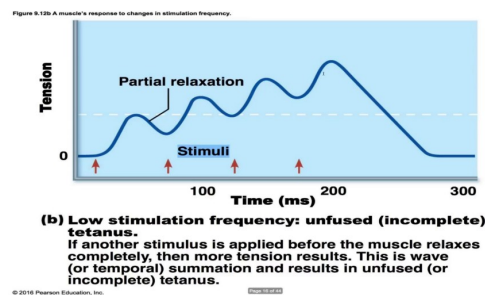


Figure 19: 2-: Anat (15:1)

© Original lecture provider; slide from LPM Dataset [27].

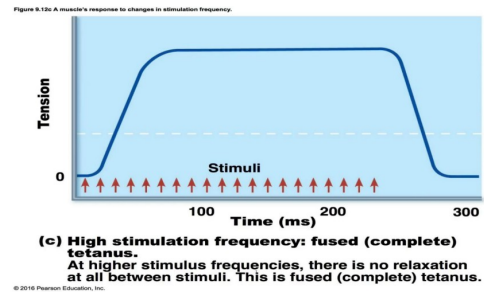


Figure 20: 3-: Anat (18:1)

© Original lecture provider; slide from LPM Dataset [27].

Let's start unimodal – Dynamic Time Warping

- We have two unaligned temporal unimodal signals
 - $X = [x_1, x_2, \dots, x_{n_x}] \in \mathbb{R}^{d \times n_x}$
 - $Y = [y_1, y_2, \dots, y_{n_y}] \in \mathbb{R}^{d \times n_y}$
- Find set of indices to minimize the alignment difference:

$$L(p_t^x, p_t^y) = \sum_{t=1}^T \|x_{p_t^x} - y_{p_t^y}\|_2^2$$

Figure 21: 4-: ML (8:1)

© Original lecture provider; slide from LPM Dataset [27].

Stephan Curve (FA)

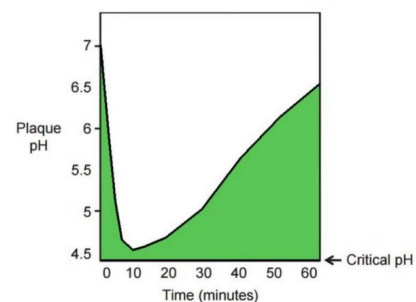


Figure 22: 5-: Dental (10:1)

© Original lecture provider; slide from LPM Dataset [27].

Diagram



Figure 23: 1+: Dental (12:1)

© Original lecture provider; slide from LPM Dataset [27].

Graph Neural Nets

Key idea: Generate node embeddings based on local neighborhoods in a recursive manner

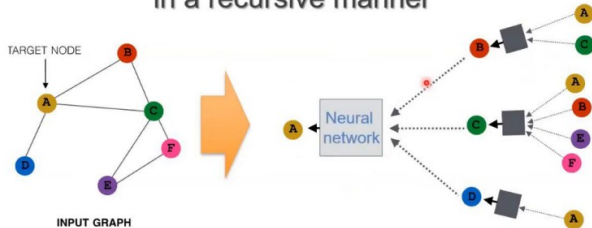


Figure 24: 2+: ML (21:1)

© Original lecture provider; slide from LPM Dataset [27].

Structure and Function: Arteries

Structure

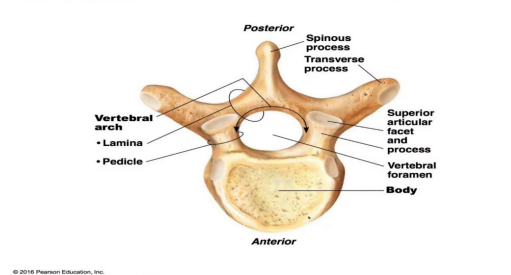
- Thick-walled
- Three layers



Figure 25: 3+: Bio (18:1)

© Original lecture provider; slide from LPM Dataset [27].

Figure 7.18 Typical vertebral structures.



© 2018 Pearson Education, Inc.

Figure 26: 4+: Anat (18:1)

© Original lecture provider; slide from LPM Dataset [27].

Animation: Rotatable Spine (Horizontal)



© 2018 Pearson Education, Inc.

Figure 27: 5+: Anat (20:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 7: ALT text for Diagram Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	Two illustrations demonstrating infrabulge clasp designs, which originate from below the height of contour. The top illustration shows a T-Bar clasp, where a metal bar extends upward from the red gingival tissue and splits into a horizontal 'T' shape, with both ends of the 'T' contacting the tooth surface. The bottom illustration shows a modified T-bar clasp, where the metal bar extends upward from the gingiva and curves to contact the tooth with only a single horizontal extension, resembling a lowercase 'r'.
1+	N	Two diagrams illustrate a T-Bar clasp engaging an abutment tooth next to an edentulous space. In both the top and bottom views, a metal bar originates from the pink gingival area below the tooth and extends upward, splitting into a T-shape. The horizontal arms of the 'T' contact the buccal surface of the tooth near the gingival margin.
2+	Y	A diagram illustrating how node embeddings are generated based on local neighborhoods in a recursive manner. On the left is the 'INPUT GRAPH' with nodes A through F. An arrow points to a 'TARGET NODE' A. A large orange arrow points to the right, showing the recursive aggregation process for node A. Node A is connected to a 'Neural network' box, which receives input from its neighbors B, C, and D. Each of these neighbors (B, C, D) is connected to its own gray box, which in turn receives input from their respective neighbors (A, C for B; A, B, E, F for C; A for D).
2+	N	A diagram showing how node embeddings are generated based on local neighborhoods. On the left is the 'INPUT GRAPH' with nodes A through F. Node A is highlighted as the 'TARGET NODE'. A large orange arrow points to the right, where a computational graph is shown. The target node A is at the center, receiving input from a 'Neural network' block. This block aggregates information from A's neighbors: B, C, and D. Each of these neighbors (B, C, D) is shown receiving input from their own neighbors through grey boxes representing neural network layers. For example, B receives input from A and C; C receives input from A, B, E, and F; D receives input from A.
3+	Y	A 3D cross-section diagram of an artery illustrating its thick-walled structure. The artery is composed of three distinct layers: an inner lining called the Tunica intima, a thick middle muscular layer called the Tunica media, and an outer protective layer called the Tunica externa. The central empty space where blood flows is the lumen.
3+	N	A 3D cross-section diagram of an artery showing its thick-walled, three-layer structure. The outermost layer is a tough protective sheath, the middle layer is thick smooth muscle, and the innermost layer is a smooth lining surrounding the hollow central lumen.
4+	Y	A superior view of a typical vertebra. Labels point to the posterior aspect, spinous process, transverse process, superior articular facet and process, vertebral foramen, body, anterior aspect, vertebral arch (composed of lamina and pedicle).
4+	N	A superior view of a typical vertebra highlights its main structural components. The anterior portion is the large, solid body. Posterior to the body is the vertebral arch, which encloses the central vertebral foramen. The vertebral arch consists of two pedicles connecting to the body and two laminae that meet at the posterior midline. Projecting from the arch are the posterior spinous process, two lateral transverse processes, and two superior articular processes with facets.
5+	Y	A photograph of a rotatable spine model in a horizontal position, showing a single vertebra from a superior view.
5+	N	A superior view of a single vertebra against a black background, showing the thick anterior body, the central vertebral foramen, and the posterior spinous and transverse processes.

Table 8: ALT text for Diagram Worst Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1-	Y	Two examples of implicit multimodal alignment. The first is a diagram showing image-sentence similarity, where an image is broken down into 'RCNN detections' (image fragments) and a sentence is broken into 'Dependency Relations' (sentence fragments). A central matrix shows the inner product similarity between these fragments, with red arrows highlighting the strongest alignments. The second example shows Visual Question Answering (VQA) with two image pairs. One asks 'What season does this appear to be?' and the other 'What is soaring in the sky?', showing the original image and a blurred version that highlights the region the model focused on to answer the question.
1-	N	A diagram illustrating implicit multimodal alignment. On the left, an image of a child and a dog is processed through RCNN detections, extracting image fragments. These fragments are compared with sentence fragments derived from dependency relations of the sentence 'Black dog chasing a young child'. A matrix shows the inner product fragment similarity between the image and sentence fragments. On the right, two examples of visual question answering are shown. The first asks 'What season does this appear to be?' with the ground truth (GT) and model answer both being 'fall'. The second asks 'What is soaring in the sky?' with GT and model answer both being 'kite'. Both examples show the original image and a blurred version highlighting the areas the model focused on to answer the question.
2-	Y	A diagram showing the coordinate-descent optimization process for Canonical Time Warping. A curved blue arrow labeled 'Generalized Eigen-decomposition' points from the temporal alignment parameters (W_x, W_y) to the cross-modal spatial alignment parameters (U, V). A red arrow points back, illustrating the iterative process of fixing one set of parameters to optimize the other.
2-	N	A diagram illustrating the optimization process for Canonical Time Warping. The equation for the objective function $L(U, V, W_x, W_y)$ is shown at the top. Below it, text explains that it is optimized by Coordinate-descent: fix one set of parameters, optimize another. A curved blue arrow labeled 'Generalized Eigen-decomposition' points back and forth between two sets of parameters: ' W_x, W_y ' (temporal alignment) and ' U, V ' (spatial alignment), indicating the iterative nature of the optimization process.
3-	Y	A diagram illustrating human visual attention. An eye is shown looking at the text 'Once upon a time there was a little fox'. Concentric yellow circles overlay the text, representing different fields of vision. The innermost circle is labeled 'Foveal' (high resolution, 2 degrees), surrounded by 'Parafoveal' (2-5 degrees), and the outermost is 'Peripheral' (5+ degrees). Red dashed lines connect the eye to the foveal region, showing where attention is sharply focused.
3-	N	A diagram illustrating attention in humans. On the left, an eye is shown looking towards the right. Red dashed lines project from the eye to a series of concentric circles on the right. The innermost circle is labeled '2-5°' and 'Foveal', representing the area of high-resolution vision. The text 'time there was a little fox liv' is superimposed on the circles, with the word 'fox' falling within the foveal area, appearing sharp and clear. The surrounding circles are labeled 'Parfoveal' and 'Peripheral', representing areas of lower resolution vision where the text appears blurry. This demonstrates the concept of foveal vision, where humans only see in high resolution in a small 2-degree area.
4-	Y	Illustration titled 'Figure 19.5d Gross anatomy of the heart, Posterior surface view.' The back of the heart is shown, dominated by the left atrium and left ventricle. Four pulmonary veins enter the left atrium, bringing oxygenated blood from the lungs. The large coronary sinus runs horizontally along the posterior coronary sulcus, collecting deoxygenated blood from cardiac veins and emptying into the right atrium. The posterior interventricular artery and middle cardiac vein run down the posterior interventricular sulcus between the ventricles. The aortic arch and pulmonary arteries are visible at the top.
4-	N	A posterior surface view of the gross anatomy of the heart. The left atrium, left ventricle, right atrium, and right ventricle are visible. Major vessels shown include the aorta, superior and inferior vena cava, and right and left pulmonary arteries and veins. The coronary sinus is prominent in the coronary sulcus, receiving blood from the great, middle, and small cardiac veins. The posterior interventricular artery is shown in the posterior interventricular sulcus. A yellow circle highlights the circumflex artery continuing to the posterior side.
5-	Y	Illustration of a frontal section of the heart demonstrating the directional flow of blood. Blue arrows show oxygen-poor blood entering the right atrium from the venae cavae, flowing down through the tricuspid valve into the right ventricle, and being pumped up through the pulmonary valve into the pulmonary trunk and out the pulmonary arteries. Red arrows show oxygen-rich blood entering the left atrium from the pulmonary veins, flowing down through the mitral valve into the left ventricle, and being pumped up through the aortic valve into the aorta and its branches.
5-	N	A schematic diagram of the heart showing the direction of blood flow. Blue arrows indicate oxygen-poor blood entering the right atrium from the venae cavae, flowing into the right ventricle, and exiting through the pulmonary trunk to the pulmonary arteries. Red arrows indicate oxygen-rich blood entering the left atrium from the pulmonary veins, flowing into the left ventricle, and exiting through the aorta to the body. A yellow circle highlights the left pulmonary artery.

Implicit multimodal-alignment

- Implicit alignment - uses internal latent alignment of modalities in order to better solve various problems

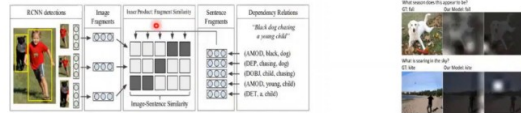


Figure 28: 1-: ML (5:1)

© Original lecture provider; slide from LPM Dataset [27].

Canonical Time Warping

$$L(\mathbf{U}, \mathbf{V}, \mathbf{W}_x, \mathbf{W}_y) = \|\mathbf{U}^T \mathbf{X} \mathbf{W}_x - \mathbf{V}^T \mathbf{Y} \mathbf{W}_y\|_F^2$$

Optimized by Coordinate-descent – fix one set of parameters, optimize another

Generalized Eigen-decomposition

$$\mathbf{W}_x, \mathbf{W}_y \quad \mathbf{U}, \mathbf{V}$$

[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Torre, 2009, NIPS]

Figure 29: 2-: ML (17:1)

© Original lecture provider; slide from LPM Dataset [27].

Attention in humans

- Foveal vision – we only see in “high resolution” in 2 degrees of vision

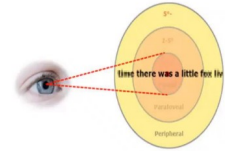
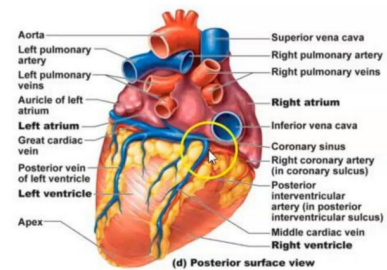


Figure 30: 3-: ML (25:1)

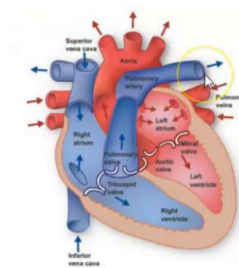
© Original lecture provider; slide from LPM Dataset [27].



© 2014 Pearson Education, Inc.

Figure 31: 4-: Bio (19:1)

© Original lecture provider; slide from LPM Dataset [27].



© 2014 Pearson Education, Inc.

Figure 32: 5-: Bio (40:1)

© Original lecture provider; slide from LPM Dataset [27].

Composite

Circumferential (Akers) Clasp

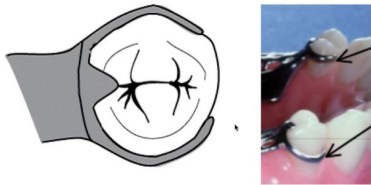


Figure 33: 1+: Dental (7:1)

© Original lecture provider; slide from LPM Dataset [27].



Figure 34: 2+: Dental (26:1)

© Original lecture provider; slide from LPM Dataset [27].



Figure 35: 3+: ML (59:1)

© Original lecture provider; slide from LPM Dataset [27].

Our food is made up of matter, and

Everything that has **matter** can be “boiled” down to different combinations of **elements**

More than 1300 chemicals in a cup of coffee!

Caffeine
 $C_8H_{10}N_4O_2$

The slide features a cup of coffee on the left, a molecular model of caffeine in the center, and the chemical formula $C_8H_{10}N_4O_2$ on the right. A blue arrow points from the text "More than 1300 chemicals in a cup of coffee!" to the molecular model. The slide has a blue footer bar.

Figure 36: 4+: Bio (3:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 9: ALT text for Composite Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	A composite visual showing a Circumferential or Akers clasp. On the left, a line drawing from an occlusal view shows the clasp assembly where both the retentive and reciprocal arms originate from a single minor connector and rest area, wrapping around opposite sides of the tooth. On the right, a photograph of a partial denture on a typodont shows the metal clasp arms in practice, with two black arrows pointing to the clasps as they encircle the buccal surfaces of two adjacent premolars.
1+	N	A composite visual shows a diagram and a photograph of a Circumferential, or Akers, clasp. On the left, a top-down diagram shows the metal clasp assembly originating from a minor connector, with two arms encircling the buccal and lingual surfaces of a molar. On the right, a photograph shows a partial denture on a dental model, with black arrows pointing to the metal clasp arms wrapping around the suprabulge portion of the abutment teeth.
2+	Y	Photograph of a double-ended enamel hatchet with a close-up view of its working end, showing a blade with a single bevel used for planing walls of enamel. The formula '10-7-14' is displayed in red.
2+	N	A composite graphic of an enamel hatchet. The main image shows a double-ended metal instrument with angled shanks and flat, single-beveled cutting blades aligned with the handle. Below is a diagram showing a close-up profile of the working end, highlighting the flat blade and beveled cutting edge. The red text '10-7-14' represents the instrument's formula.
3+	Y	A grid displays image translation results generated by a Cycle GAN. The top two rows show translation from daytime landscape photos in the 'Input' column to winter or snowy landscapes in the 'Generated samples' columns, compared against a 'Ground truth' reference. The bottom two rows show translation from line sketches of shoes in the 'Input' column to fully colored and textured shoe photos in the 'Generated samples' columns, again compared to a 'Ground truth' photo.
3+	N	Examples of image-to-image translation. The top two rows show translation between day and night scenes of a city. The bottom two rows show translation from edge sketches of shoes to photorealistic images of shoes in various colors and styles. Columns show the Input, Ground truth, and several Generated samples.
4+	Y	Composite graphic showing a cup of coffee next to a ball-and-stick model of a caffeine molecule. Text indicates there are more than 1300 chemicals in a cup of coffee. The caffeine molecule model shows its chemical formula, C ₈ H ₁₀ N ₄ O ₂ , with black spheres representing carbon, white spheres for hydrogen, blue spheres for nitrogen, and red spheres for oxygen.
4+	N	A composite graphic showing a cup of coffee next to coffee beans, linked by a bracket to the text 'More than 1300 chemicals in a cup of coffee!'. An arrow points from this text to a 3D ball-and-stick molecular model of caffeine, labeled 'Caffeine C ₈ H ₁₀ N ₄ O ₂ '.
5+	Y	Collage of four images representing different types of bonds. Top left shows a person diving into a swimming pool. Top middle shows a drawing of two puzzle pieces fitting together with the text 'You complete me'. Right shows two hands tying a knot with multiple colorful strings. Bottom shows a cartoon of a large football player and a small child pulling on opposite ends of a rope.
5+	N	A person belly-flopping into a swimming pool. A drawing of two puzzle pieces fitting together with a heart above them and the text 'You complete me.' below. Two hands holding a complex knot made of colorful strings. A cartoon of a small child and a large football player playing tug-of-war.

Table 10: ALT text for Composite Worst Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1-	Y	A composite visual showing the four valves of the heart. In the center is a cross-section diagram of the heart with arrows indicating the forward flow of blood through the chambers and valves. A yellow circle highlights the right atrioventricular valve between the right atrium and right ventricle. To the left, an arrow points to a photograph of an open heart valve, with its string-like chordae tendineae circled in red. To the right, an arrow points to a photograph of a closed valve viewed from above, displaying its cusps tightly sealed together. Handwritten red letters 'R' and 'L' denote the right and left sides of the visual.
1-	N	A central cross-section diagram of the heart highlights its internal valves. An arrow points from the right side of the heart to an inset photograph on the left, labeled 'R', showing the string-like chordae tendineae attached to a valve. Another arrow points from the left side of the heart to an inset photograph on the right, labeled 'L', showing a close-up of a closed, three-flap valve resembling a doorway.
2-	Y	Three circular graphics accompany the learning objectives. The first is a flowchart showing the levels of biological organization progressing from atom to organism. The second is an anatomical illustration of a human torso highlighting the digestive system organs. The third is a detailed illustration of the coiled small and large intestines. A green checkmark indicates the first learning objective is complete.
2-	N	Three circular graphics illustrating biological organization. The top left circle shows a flowchart of the 'Levels of Organization' progressing from Atom to Organism, with 'Cell' highlighted in yellow. The middle circle shows an illustration of a human torso with the digestive system organs visible inside. The bottom left circle shows a detailed illustration of the human digestive tract. A large green checkmark is placed next to the first learning objective text on the right.
3-	Y	An anatomical diagram of the side of a human face highlights the three main salivary glands: the parotid gland in front of the ear, the submandibular gland below the jaw, and the sublingual gland under the tongue. A humorous inset image shows a slice of pizza floating in a bathtub, illustrating the concept of food being bathed in saliva for chemical digestion. Text notes that saliva contains amylase, an enzyme that breaks down starch.
3-	N	On the left, an anatomical diagram of a human head in profile highlights the salivary glands in pink, specifically labeling the 'Parotid gland' and 'Submandibular gland'. On the top right, a circular image shows a slice of pizza floating in a white bathtub filled with water, serving as a visual analogy for the pizza being bathed in saliva.
4-	Y	Photograph of a double-ended scaler and diagrams of four Gracey scaler tips. Red text indicates 'Universal= anywhere' and 'Gracey= specific area', explaining that universal scalers can be used anywhere in the mouth, while Gracey scalers are designed for specific teeth and areas.
4-	N	A composite graphic showing a dental scaler. On the left is a photograph of a double-ended metal scaler with sharply curved, pointed working ends. On the right are four diagrams of specific scaler working ends labeled 'GRACEY 1-2', 'GRACEY 7-8', 'GRACEY 11-12', and 'GRACEY 13-14'. Above the scaler, red text reads 'Universal= anywhere' and 'Gracey= specific area', indicating their intended usage.
5-	Y	Photograph of a double-ended discoid-cleoid carver. Close-up views show the two distinct working ends: the claw-like cleoid end used for carving grooves, and the circular discoid end used for carving pits and fossae into amalgam.
5-	N	A composite graphic of a discoid-cleoid carver. The main image shows a double-ended metal instrument. To the left are two close-up photographs of the working ends within circles: one end is circular and disc-shaped (discoid), and the other is pointed and spade-shaped (cleoid). Below is a diagram showing a side profile of the pointed cleoid end.

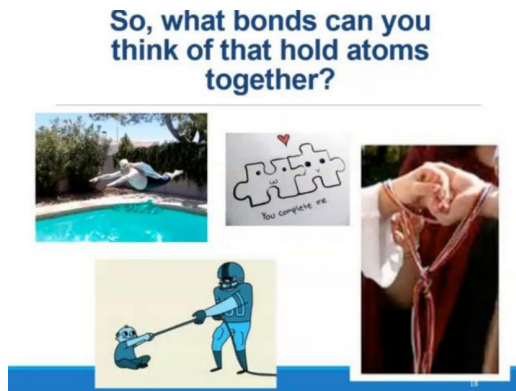


Figure 37: 5+: Bio (18:1)

© Original lecture provider; slide from LPM Dataset [27].

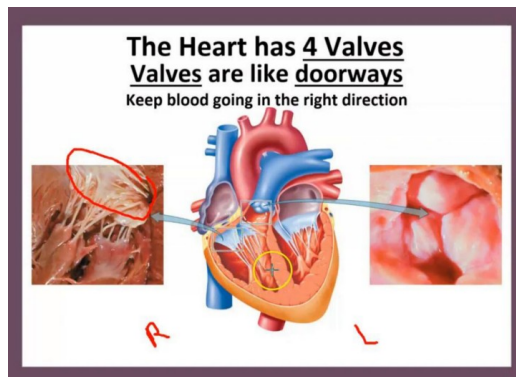


Figure 38: 1-: Bio (10:1)

© Original lecture provider; slide from LPM Dataset [27].

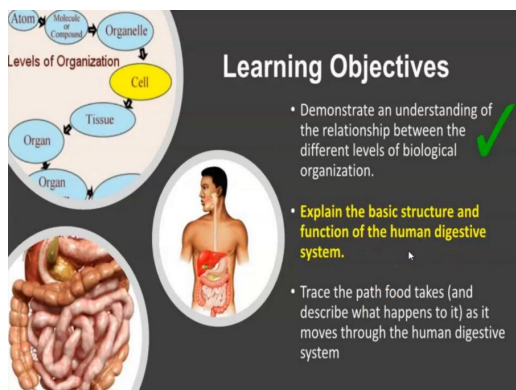


Figure 39: 2-: Bio (8:1)

© Original lecture provider; slide from LPM Dataset [27].

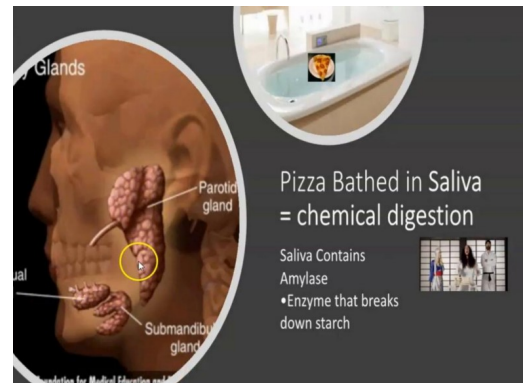


Figure 40: 3-: Bio (18:1)

© Original lecture provider; slide from LPM Dataset [27].

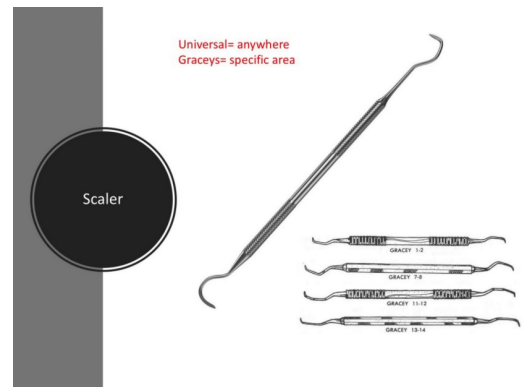


Figure 41: 4-: Dental (22:1)

© Original lecture provider; slide from LPM Dataset [27].

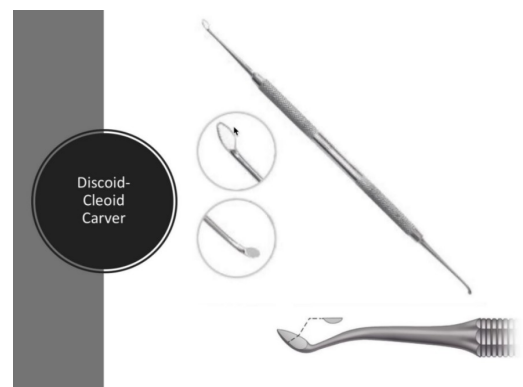


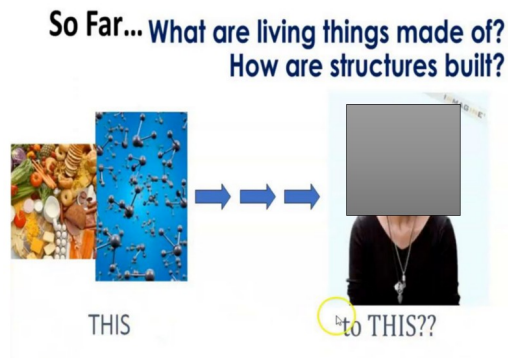
Figure 42: 5-: Dental (34:1)

© Original lecture provider; slide from LPM Dataset [27].

Infographic

Top Best ALT Text Examples. No examples matched this direction (ie. there were no examples with a positive transcript-minus-no-transcript difference in this group).

Top Worst ALT Text Example. Only 1 example is available for this direction (ie. there was only 1 example with a negative transcript-minus-no-transcript difference in this group).



(a) Slide

A sequence of three images connected by right-pointing arrows illustrates the building blocks of life. The sequence begins with a diverse pile of food, including vegetables, grains, and proteins. An arrow points to a 3D molecular model of interconnected atoms, representing macromolecules. Another arrow points to a photograph of a woman, posing the question of how food and molecules build a living human structure.

(b) ALT Text with transcript

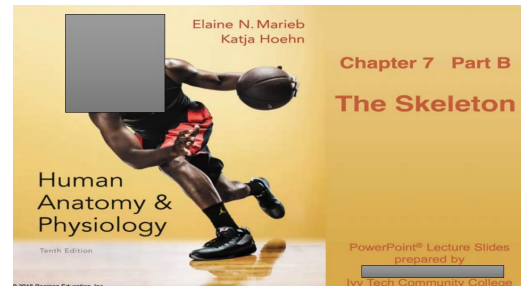
A sequence of three images connected by three right-pointing blue arrows, illustrating the building blocks of life. The first image shows a pile of various foods including vegetables, bread, and cheese. The second image shows a 3D molecular model with dark spheres connected by gray rods against a blue background. The arrows point from these molecular structures to the third image, a photograph of a smiling woman. The text 'THIS' is below the food and molecules, and 'to THIS??' is below the woman.

(c) ALT Text without transcript

Figure 43: A Biology slide (slide 6, visual 1) with the first worst **Infographic** ALT text example given transcript context. Note that a human individual's face on the slide has been manually blocked out by our team for privacy.

© Original lecture provider; slide from LPM Dataset [27].

Photograph



(a) 1+: Anat (1:1). Note that a human individual's face on the slide and a name has been manually blocked out by our team for privacy.

Superhuman results on Montezuma's Revenge



(b) 2+: ML (52:1)

Ring Clasp

- Used when undercut is adjacent to BES



(c) 3+: Dental (10:1)

Figure 44: 1+, 2+ and 3+ best visuals for Photograph (slide #: visual #). For 1+, note that a human individual's face on the slide and a name has been manually blocked out by our team for privacy.

© Original lecture provider; slides from LPM Dataset [27].

Embrasure Clasp



Figure 45: 4+: Dental (11:1)

© Original lecture provider; slide from LPM Dataset [27].

Direct Retainer (Clasp Assembly)

- Extracoronary retainer
 - More common, conventional “clasp” design
 - Clasps should encircle a tooth **at least 180°**
- Intracoronary retainer
 - A precision attachment with key and keyway pattern
 - More esthetic because no clasps



Figure 47: 1-: Dental (3:1)

© Original lecture provider; slide from LPM Dataset [27].

11 Friendships In Adulthood

Chapter

- Adult friendships develop over several stages
 - Acquaintanceship
 - Buildup
 - Continuation
 - Deterioration
 - Ending



Figure 46: 5+: Psych (3:1). Note that human individuals' faces on the slide have been manually blocked out by our team for privacy.

© Original lecture provider; slide from LPM Dataset [27].

Learning Objectives

- What are the organs of the cardiovascular system, and how does blood flow through them?
- Explain why different blood vessels are built the way they are?
- Solve case studies related to coronary cardiovascular disease.
- Describe how blood pressure relates to the cardiac cycle.

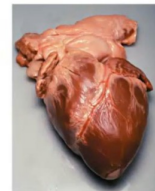


Figure 48: 2-: Bio (3:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 11: ALT text for Photograph Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	A photograph of a basketball player dribbling a basketball. The player is wearing a black tank top, black shorts with red accents, black socks, and black sneakers. The background is a solid yellow color. The text on the slide reads 'Elaine N. Marieb, Katja Hoehn, Chapter 7 Part B, The Skeleton, Human Anatomy & Physiology, Tenth Edition, PowerPoint Lecture Slides prepared by [redacted for privacy], Ivy Tech Community College'.
1+	N	A male basketball player in a black jersey and red and black shorts is shown mid-stride, dribbling a basketball with his left hand. He is looking forward with intense focus, set against a solid yellow background.
2+	Y	A screenshot of the Atari game Montezuma's Revenge. The player character is standing on a platform above a pit, with a key visible on the left and a skull enemy below. Text next to the image explains that maximizing curiosity encourages the agent to explore environments with sparse rewards, stochasticity, and unknown dynamics.
2+	N	A screenshot of the Atari game Montezuma's Revenge, showing the player character in a room with platforms, ladders, and a skull enemy.
3+	Y	A black-and-white photograph showing a ring clasp on an isolated molar, marked with red annotations to indicate undercut locations. A red circle with an 'X' is drawn on the buccal surface opposite the edentulous space, indicating an inadequate undercut for a standard clasp. Two red circles highlight the mesial areas adjacent to the edentulous space where the usable undercut is located. The metal ring clasp originates from the mesial framework, wraps entirely around the lingual, distal, and buccal surfaces of the tooth, and dives down to engage the mesial undercut.
3+	N	A close-up, black-and-white photograph shows a ring clasp on a dental model, marked with red annotations. The metal clasp nearly completely encircles the isolated abutment tooth. Red circles highlight specific contact points or rests on the occlusal and lateral surfaces, while a red 'X' marks a section of the clasp arm on the lower left side of the tooth.
4+	Y	A photograph showing an embrasure clasp on a typodont. The metal framework extends upward between two adjacent premolars, crosses over the occlusal embrasure, and splits into two separate clasp arms. A black arrow points to this junction. The two arms then wrap around the buccal surfaces of the two premolars in opposite directions, effectively functioning as two joined circumferential clasps.
4+	N	A close-up photograph displays an embrasure clasp on a dental model. The metal framework crosses the occlusal embrasure between two adjacent posterior teeth and splits into two separate clasp arms. These arms extend in opposite directions, wrapping around the buccal surfaces of both teeth. Black arrows point to the clasp arms engaging the teeth.
5+	Y	Four young adults are walking outdoors together on a sunny day, smiling and laughing while eating ice cream cones. The group consists of two men and two women, illustrating a casual social interaction or friendship buildup. A small, simple hand-drawn stick figure is also visible in the bottom left corner of the slide.
5+	N	Four young adults walk together outdoors on a sunny day. They are smiling, laughing, and eating ice cream cones, representing a happy, communal friendship.

Table 12: ALT text for Photograph Worst Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1-	Y	A photograph showing an intracoronar retainer precision attachment against a blue background. The attachment consists of two small, interlocking metal pieces that fit together using a key and keyway pattern.
1-	N	A photograph displays two small, metallic precision attachment components for an intracoronar retainer, resting on a bright blue background. One component features a cylindrical slot, representing the keyway. The other component has a matching cylindrical projection, representing the key, designed to interlock with the keyway.
2-	Y	A photograph of a human heart resting on a flat surface, showing its muscular structure and surface blood vessels.
2-	N	An anatomical photograph of a human heart showing its muscular exterior, blood vessels on the surface, and upper chambers. A decorative blue heart icon is also present in the top right corner.
3-	Y	A composite visual of two photographs distinguishing infected and affected dentin. The left photo shows a cross-section of an extracted tooth with a deep carious lesion, revealing a superficial, lighter, mushy area and a deeper, darker brown, leathery area. The right photo is a close-up of a dental hand instrument scraping soft, brownish, mushy material from a cavity preparation. Blue arrows connect the red text 'Infected dentin' to the superficial mushy material in both photos, indicating it is necrotic and easily removed. Another blue arrow connects the red text 'Affected dentin' to the deeper, darker brown area in the first photo, indicating it is demineralized but not invaded by bacteria.
3-	N	Two photographs showing the progress of caries lesions in a tooth. The left image is a cross-section of an extracted tooth with a large cavity. A light blue arrow points from the text 'Infected dentin' to the superficial, darker, mushy-looking decayed area at the top of the cavity. Another light blue arrow points from the text 'Affected dentin' to the deeper, lighter-brown, harder-looking dentin layer below the infected zone. The right image shows a close-up clinical view of a cavity preparation inside a mouth, with a dental instrument scraping away soft, brown infected dentin, which is also indicated by a light blue arrow from the 'Infected dentin' text.
4-	Y	A composite of three side-by-side photographs showing different fathers affectionately holding and caring for their infant children. The images illustrate the parental role, specifically highlighting men spending time caring for their young children as they get older.
4-	N	A composite of three side-by-side photographs showing different fathers affectionately holding and caring for their infants. The left image shows a young father holding a baby who is looking at the camera. The center image shows a father smiling down at a sleeping newborn in his arms. The right image shows an older father smiling warmly while holding a baby against his chest.
5-	Y	Photograph showing two dental explorers used to provide tactile sensitivity to detect the integrity of fillings and crown margins. One is a double-ended instrument with curved, pointed ends. The other shows a single end with a different curved shape.
5-	N	A photograph showing two dental explorers. The main instrument is a double-ended metal tool with a textured handle; one end has a long, curved, hook-like tip, and the other end has a shorter, sharply angled tip. In the upper right corner, a close-up shows the working end of another explorer with a sharply angled, thin, pointed tip.

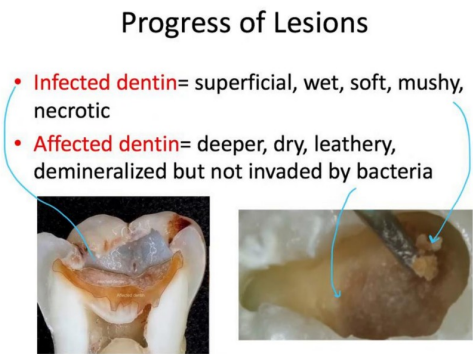


Figure 49: 3--: Dental (14:1)

© Original lecture provider; slide from LPM Dataset [27].

11 The Parental Role

Chapter

- Couples are having fewer children and waiting longer to have them and older parents are more at ease, affectionate, sensitive, and supportive
- Men who become fathers in their 30s spend up to 3 times as much time caring for their preschool children

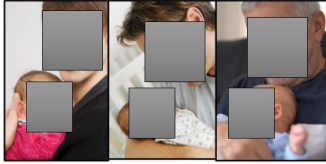


Figure 50: 4-: Psych (9:1). Note that human individuals' faces on the slide have been manually blocked out by our team for privacy.

© Original lecture provider; slide from LPM Dataset [27].

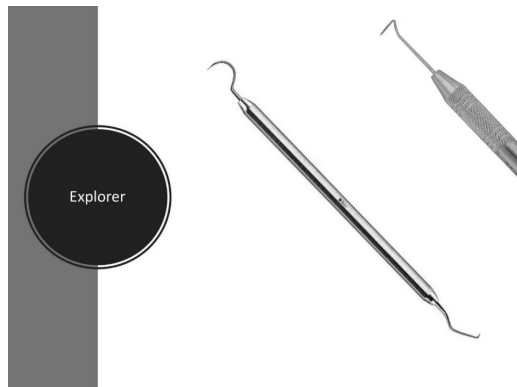


Figure 51: 5-: Dental (9:1)

© Original lecture provider; slide from LPM Dataset [27].

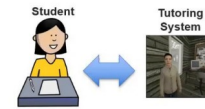
Table

Top 4 Best ALT Text Examples. Only 4 examples are available for this direction (ie. there were only 4 examples with a positive transcript-minus-no-transcript difference in this group).

Top Worst ALT Text Examples. No examples matched this direction (ie. there were no examples with a negative transcript-minus-no-transcript difference in this group).

Example: Inferring Emotion from Interaction Logs

[Sabourin et al., 2011]



	Emotion Accuracy	Valence Accuracy
Baseline	22.4%	54.5%
Naïve Bayes	18.1%	51.2%
Bayes Net	25.5%	66.8%
Dynamic BN	32.6%	72.6%

Language Technologies Institute

Carnegie Mellon University

Figure 52: 1+: ML (17:1)

© Original lecture provider; slide from LPM Dataset [27].

Experience replay

- To remove correlations, build data-set from agent's own experience

S_1, a_1, r_2, S_2
S_2, a_2, r_3, S_3
S_3, a_3, r_4, S_4
...
$S_t, a_t, r_{t+1}, S_{t+1}$

→ S, a, r, S'

exploration, epsilon greedy is important!

- Sample random mini-batch of transitions (s,a,r,s') from D

Figure 53: 2+: ML (46:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 22.1 The Upper Respiratory System

STRUCTURE	DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES	FUNCTION
Nose (external nose and nasal cavity)	Jutting external portion is supported by bone and cartilage. Internal nasal cavity is divided by median nasal septum and lined with mucosa. Roof of nasal cavity contains olfactory epithelium.	Produces mucus, filters, warms, and moistens incoming air; resonance chamber for speech. Receptors for sense of smell.
Paranasal sinuses	Mucosa-lined, air-filled cavities in cranial bones surrounding nasal cavity.	Lighten skull; also may warm, moisten, and filter incoming air.
Pharynx	Passageway connecting nasal cavity to larynx and oral cavity to esophagus. Three subdivisions: nasopharynx, oropharynx, and laryngopharynx. Houses tonsils (lymphoid tissue masses involved in protection against pathogens).	Passageway for air and food. Facilitates exposure of immune system to inhaled antigens.

© 2014 Pearson Education, Inc.

© 2014 Pearson Education, Inc.

Figure 54: 3+: Anat (6:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 13: ALT text for Table Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	A table compares the performance of four models on emotion and valence accuracy for inferring student states from interaction logs. The models are Baseline, Naïve Bayes, Bayes Net, and Dynamic BN. The Dynamic BN model achieves the highest accuracy in both categories, scoring 32.6% for Emotion Accuracy and 72.6% for Valence Accuracy. The Baseline model scores 22.4% and 54.5%, Naïve Bayes scores 18.1% and 51.2%, and Bayes Net scores 25.5% and 66.8% respectively. Above the table is the illustration of the student interacting with the tutoring system.
1+	N	A table comparing the accuracy of different models for inferring emotion and valence. The models are Baseline, Naïve Bayes, Bayes Net, and Dynamic BN. The Emotion Accuracy scores are 22.4%, 18.1%, 25.5%, and 32.6% respectively. The Valence Accuracy scores are 54.5%, 51.2%, 66.8%, and 72.6% respectively. The Dynamic BN model has the highest accuracy for both.
2+	Y	A diagram illustrating experience replay. A table shows a sequence of experiences: (s_1, a_1, r_2, s_2) , (s_2, a_2, r_3, s_3) , etc. An arrow points from the table to a generic tuple ' s, a, r, s' ', indicating that random mini-batches of these transitions are sampled from the dataset.
2+	N	A table representing an experience replay buffer. It contains rows of transitions: (s_1, a_1, r_2, s_2) , (s_2, a_2, r_3, s_3) , (s_3, a_3, r_4, s_4) , ..., $(s_t, a_t, r_{t+1}, s_{t+1})$. An arrow points from the table to a generic transition tuple ' s, a, r, s' '.
3+	Y	A table titled 'Table 22.1 The Upper Respiratory System'. It has three columns: 'STRUCTURE', 'DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES', and 'FUNCTION'. The rows detail the Nose (external nose and nasal cavity), Paranasal sinuses, and Pharynx. The Nose is described as jutting external portion supported by bone and cartilage, with an internal nasal cavity divided by a midline nasal septum and lined with mucosa. Its functions include producing mucus, filtering, warming, and moistening incoming air, serving as a resonance chamber for speech, and housing receptors for the sense of smell. Paranasal sinuses are mucosa-lined, air-filled cavities in cranial bones surrounding the nasal cavity, functioning to lighten the skull and warm, moisten, and filter incoming air. The Pharynx is a passageway connecting the nasal cavity to the larynx and oral cavity to the esophagus, with three subdivisions: nasopharynx, oropharynx, and laryngopharynx. It functions as a passageway for air and food and houses tonsils to facilitate exposure of the immune system to inhaled antigens.
3+	N	A table titled 'Table 22.1 The Upper Respiratory System'. It has three columns: 'STRUCTURE', 'DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES', and 'FUNCTION'. The rows detail the Nose (external nose and nasal cavity), Paranasal sinuses, and Pharynx, providing their descriptions and functions.
4+	Y	A table titled 'Table 22.2 The Lower Respiratory System'. It has three columns: 'STRUCTURE', 'DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES', and 'FUNCTION'. The rows detail the Larynx, Trachea, Bronchial tree, Alveoli, Lungs, and Pleurae. The Larynx connects the pharynx to the trachea, has a framework of cartilage and dense connective tissue, and houses vocal folds. The Trachea is a flexible tube running from the larynx and dividing inferiorly into two main bronchi. The Bronchial tree consists of right and left main bronchi, which subdivide to form lobar and segmental bronchi and bronchioles. The Alveoli are microscopic chambers at the termini of the bronchial tree, serving as the main sites of gas exchange. The Pleurae are serous membranes that produce lubricating fluid and compartmentalize the lungs.
4+	N	A table titled 'Table 22.2 The Lower Respiratory System'. It has three columns: 'STRUCTURE', 'DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES', and 'FUNCTION'. The rows detail the Larynx, Trachea, Bronchial tree, Alveoli, Lungs, and Pleurae, providing their descriptions and functions.

Table 22.2 The Lower Respiratory System

STRUCTURE	DESCRIPTION, GENERAL AND DISTINCTIVE FEATURES	FUNCTION
Larynx	Connects pharynx to trachea. Has framework of cartilage and dense connective tissue. Opening (glottis) can be closed by epiglottis or vocal folds. Houses vocal folds (true vocal cords).	Air passageway; prevents food from entering lower respiratory tract. Voice production.
Trachea	Flexible tube running from larynx and dividing inferiorly into two main bronchi. Walls contain C-shaped cartilage that are incomplete posteriorly where connected by tracheals.	Air passageway; cleans, warms, and moistens incoming air.
Bronchial tree	Consists of right and left main bronchi, which subdivide within the lungs to form lobar and segmental bronchi and bronchioles. Bronchioles walls lack cartilage but contain complete layer of smooth muscle. Constriction of this muscle impedes expiration.	Air passageways connecting trachea with alveoli; cleans, warms, and moistens incoming air.
Alveoli	Microscopic chambers at termini of bronchial tree. Walls of simple squamous epithelium overlie thin basement membrane. External surfaces are intimately associated with pulmonary capillaries. Special alveolar cells produce surfactant.	Main sites of gas exchange. Reduce surface tension; helps prevent lung collapse.
Lungs	Paired composite organs that flank mediastinum in thorax. Composed primarily of alveoli and respiratory passageways. Strong is elastic connective tissue, allowing lungs to recoil passively during expiration.	House respiratory passages smaller than the main bronchi.
Pleurae	Serous membranes. Parietal pleura lines thoracic cavity; visceral pleura covers external lung surface.	Produce lubricating fluid and compartmentalize lungs.

Figure 55: 4+: Anat (7:1)

© Original lecture provider; slide from LPM Dataset [27].

Illustration



Figure 60: 5+: Psych (21:1)

© Original lecture provider; slide from LPM Dataset [27].

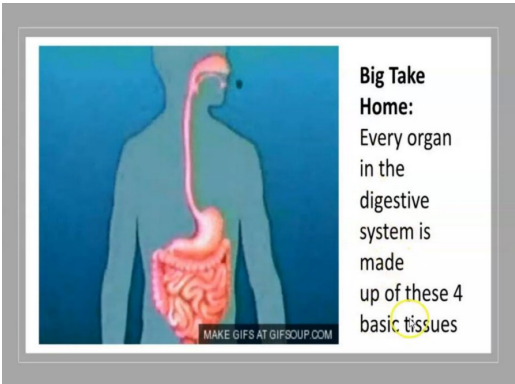


Figure 56: 1+: Bio (11:1)

© Original lecture provider; slide from LPM Dataset [27].

Top 3 Worst ALT Text Examples. Only 3 examples are available for this direction (i.e. there were only 3 examples with a negative transcript-minus-no-transcript difference in this group).



Figure 57: 2+: Bio (27:1)

© Original lecture provider; slide from LPM Dataset [27].



Figure 58: 3+: Dental (1:1)

© Original lecture provider; slide from LPM Dataset [27].

Convergence

- Amazing result: Q-learning converges to optimal policy -- even if you're acting suboptimally!
- This is called **off-policy learning**
- Caveats:
 - You have to explore enough
 - You have to eventually make the learning rate small enough
 - ... but not decrease it too quickly



Slides from Abbeel

Figure 59: 4+: ML (40:1)

© Original lecture provider; slide from LPM Dataset [27].



Figure 61: 1-: Psych (28:1)

© Original lecture provider; slide from LPM Dataset [27].

Policy

Definition: A policy is a distribution over actions given states

$$\pi(a|s) = \Pr(A_t = a | S_t = s), \forall t$$

- A policy fully defines the behavior of an agent
- The policy is stationary (time-independent)
- During learning, the agent changes its policy as a result of experience

Special case: deterministic policies




Figure 62: 2-: ML (9:1)

© Original lecture provider; slide from LPM Dataset [27].



Figure 63: 3-: Bio (33:1)

© Original lecture provider; slide from LPM Dataset [27].

Table 14: ALT text for Illustration Best Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1+	Y	An anatomical illustration of a human silhouette highlights the continuous path of the digestive tract in pink. The highlighted organs include the mouth, esophagus, stomach, small intestine, and large intestine, demonstrating the interconnected system.
1+	N	A blue silhouette of a human torso containing a highlighted pink and red digestive system. The highlighted organs include the mouth, esophagus, stomach, small intestine, and large intestine, showing the continuous pathway of the digestive tract.
2+	Y	A humorous cartoon depicts a pink stomach character operating a wooden stand, similar to a lemonade stand. The sign reads 'ONLY 5 cents CHYME.' The proud stomach character pours a green liquid from a pitcher into small cups, with a speech bubble declaring, 'I made this!', illustrating the stomach's role in producing chyme to pass to the small intestine.
2+	N	A cartoon drawing of a pink, anthropomorphic stomach standing behind a wooden lemonade-style stand. The sign on the stand reads 'ONLY 5¢ CHYME'. On the counter is a pitcher and two glasses filled with a green, lumpy liquid representing chyme. The stomach character has a speech bubble that says 'I made this!'.
3+	Y	Illustration of a blue brain with a white tooth in the center, representing the topic of basic dental instruments.
3+	N	A stylized illustration of a white tooth centered within a blue, cloud-like brain structure with thick black outlines.
4+	Y	A cartoon illustration of a robot with its arms raised in the air, appearing to jump or celebrate, with yellow action lines radiating from it. Identical to the illustration on slide 38.
4+	N	A cartoon illustration of a robot with its arms raised in the air, appearing to jump or celebrate, with yellow action lines radiating from it.
5+	Y	A four-panel comic strip shows a mother breaking up a fight between two children. In the first panel, she yells, THAT IS ENOUGH!! In the second, she says, We have just brought a new baby into the world. Do you want her to think that this is what life is all about? In the third, she continues, We want her to believe that people are reasonable and kind, that there are other ways of resolving disputes without violence. We want her to believe that the world is a GOOD PLACE! In the final panel, a man standing nearby dryly adds, ...At least until she starts watching TV.
5+	N	A four-panel comic strip. In the first panel, a woman yells "THAT IS ENOUGH!!" at two fighting children in a cloud of dust. In the second panel, she says to a man, "We have just brought a new baby into the world. Do you want her to think that THIS is what life is all about?" In the third panel, she continues, "We want her to believe that people are reasonable and kind, that there are other ways of resolving disputes without violence. We want her to believe that the world is a GOOD PLACE!" while the children stand quietly. In the fourth panel, the man looks thoughtfully at the baby and replies, "...At least until she starts watching TV."

Table 15: ALT text for Illustration Worst Examples. R = rank (+/- = best/worst). T = Transcript (Y/N); D = Description

R	T	D
1-	Y	A cartoon illustration of Garfield the cat sleeping soundly on his stomach atop a light blue pillow. A thought bubble above his head contains a large letter S, indicating that he is snoring or in a deep sleep.
1-	N	The text 'The End' appears above a cartoon of Garfield the cat sleeping soundly on a light blue pillow. A speech bubble above Garfield contains the letter 'S', indicating he is snoring.
2-	Y	Two illustrations demonstrating a policy in reinforcement learning. The top illustration shows a simple grid world with a robot agent at the start position (1,1). The grid contains a diamond with a +1 reward at (4,3), a fire pit with a -1 reward at (4,2), and an obstacle at (2,2). Green arrows indicate the optimal path to the diamond. The bottom illustration shows a pair of robotic hands holding a map of the same grid world. The map displays arrows in each grid cell, representing a deterministic policy that dictates the action to take in every state to reach the diamond and avoid the fire.
2-	N	A grid world game environment showing a robot starting at position (3,1). The grid contains a pit at (2,2), a diamond with a '+1' reward at (4,3), and a fire with a '-1' reward at (4,2). Below the grid, a pair of robotic hands holds a map showing a deterministic policy, indicated by arrows in each grid cell directing the robot towards the diamond while avoiding the pit and fire.
3-	Y	A 3D illustration of a cell in the late stages of division, pinching deeply in the middle to form two separate cells. Inside, clusters of purple chromosomes are being pulled toward opposite poles by a network of thin spindle fibers.
3-	N	A 3D illustration shows a dividing cell pinching deeply in the middle. Inside the two forming halves, bundles of purple chromosomes are being pulled toward opposite poles by a network of thin, golden spindle fibers radiating from star-shaped centrosomes at each end. The text 'BioFlix Mitosis' is overlaid on the image.