# KnowVis: Knowledge-Centric Visual Summarization for Video Lectures

**Yi Xu**
City University of Hong Kong
yixu96-c@my.cityu.edu.hk

**Yifan Hou**
ETH Zürich
yifan.hou.z@gmail.com

**Xiaoyu Zhang***
City University of Hong Kong
xiaoyu.zhang@cityu.edu.hk

## Abstract

Video lectures are valuable educational resources, but their dense and lengthy format often overwhelms novice learners. This difficulty stems from a fundamental pedagogical mismatch: videos deliver transient information linearly, while human learning requires constructing interconnected cognitive networks, which could induce extreme cognitive overload for novice learners lacking prior domain knowledge. Existing video summarization methods fail to address this mismatch, as they primarily produce text-heavy, linear condensations that still demand high cognitive effort. To bridge this gap, we propose KnowVis, a framework that transforms linear video lectures into pedagogically grounded visual narratives. KnowVis first extracts a detailed concept map from multimodal video content to identify important and challenging threshold concepts, then constructs structured knowledge units, and finally synthesizes engaging visual summaries. Alongside the framework, we introduce a curated dataset of 125 educational videos across 10 academic disciplines, paired with 1,079 generated visual summaries. Extensive automated evaluations and a human study demonstrate that, compared to state-of-the-art baselines, KnowVis generates more accurate and clear visuals that successfully reduce cognitive load and significantly improve student learning effectiveness and knowledge retention.[1]

## 1 Introduction

Video lectures have long proven to be a valuable and scalable medium that democratizes access to knowledge from professional instructors across the globe (Ally, 2004; Breslow et al., 2013; Seaton et al., 2014). However, sustaining active engagement during dense, lengthy video content remains a significant challenge (Ally, 2008). For novice learners unfamiliar with a topic, passively watching a continuous stream of information often leads to cognitive fatigue (Chen and Wu, 2015). Without clear visual scaffolding, students struggle to extract core concepts, perceive their interrelationships, and maintain focus, which ultimately degrades the overall learning experience (Gutiérrez-González et al., 2024; Schacter and Szpunar, 2015).

From a pedagogical perspective, this difficulty stems from a fundamental mismatch between the linear presentation of video media and the non-linear mechanics of human cognitive processing. According to Cognitive Load Theory (Sweller, 2020) and the Cognitive Theory of Multimedia Learning (Mayer and Moreno, 1998; Mayer, 2009), human learning relies on schema construction: the process of connecting new concepts to existing knowledge networks in a recursive, interconnected manner. Video lectures, however, deliver information sequentially and transiently. This "transient information effect" forces learners to simultaneously hold earlier concepts in their working memory while processing incoming information, imposing a high *extraneous* cognitive load (Wong et al., 2012). Furthermore, for novice learners, a lack of prior knowledge inherently imposes a high *intrinsic* cognitive load when they attempt to grasp complex concepts (Kalyuga, 2005). Consequently, this confluence of high intrinsic and extraneous load overwhelms working memory, actively hindering the learner's ability to meaningfully integrate new knowledge (Sweller et al., 2011).

Existing computational efforts to mitigate these learning barriers generally fall into three categories, none of which fully resolve this pedagogical mismatch. First, textual video summarization methods condense lengthy videos into shorter abstracts (Gonzalez et al., 2023; Liu et al., 2025). However, they still output linear, text-heavy representations that require significant reading effort

*Corresponding Author

[1]Our code and dataset are available at: https://github.com/yixu-cityu/KnowVis.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 51823–51840
October 24-29, 2026 ©2026 Association for Computational Linguistics

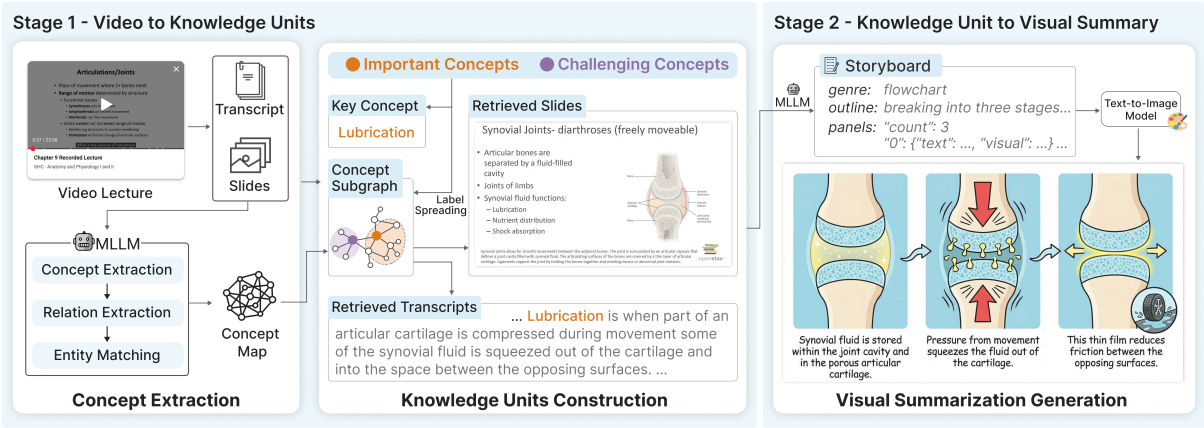


Figure 1: **Overview of the KnowVis framework.** Stage 1: It extracts a detailed *concept map* from the *video content*, identifies important and challenging concepts, and extracts concept subgraphs to construct structured *knowledge units*. Stage 2: It generates storyboards with narratives and visual descriptions to synthesize engaging *visual summaries*.

and fail to visually illustrate the underlying conceptual structures. Second, video retrieval techniques allow learners to search for and navigate to specific timestamps (Schwab et al., 2017; Zhou et al., 2025; Liu et al., 2018). However, these tools merely aid navigation without synthesizing content to make the underlying knowledge more cognitively accessible. Finally, while recent advancements in generative AI have enabled domain-specific visual generation, such as transforming math problems into diagrams (Wang et al., 2025a) or generating medical flashcards (Wu et al., 2025), these approaches rely on heavily constrained, pre-defined templates that cannot generalize to the diverse, multimodal content in open-domain educational lectures.

To bridge the gap between linear video presentation and non-linear cognitive learning, we propose KnowVis, a framework that automatically generates domain-independent, pedagogically grounded visual summaries from video lectures. Specifically, the KnowVis pipeline first processes multimodal video inputs (transcripts and slide frames) to extract a comprehensive, non-linear concept map. Guided by Threshold Concepts theory (Meek et al., 2023), KnowVis identifies key concepts that are structurally important or cognitively challenging for learners. The framework then clusters closely related concepts into knowledge units and finally synthesizes narrative visual summaries to explain them. To support this, we introduce a rigorously curated multimodal dataset comprising 125 video lectures across 10 diverse academic disciplines, paired with 1,079 KnowVis-generated visual summaries, concept graphs, and source transcripts.

We comprehensively evaluate KnowVis using both automated and human-centric metrics. Using a Large Language Model (LLM) as an automated judge (Zheng et al., 2023), we demonstrate that KnowVis outperforms state-of-the-art multimodal baselines in accuracy and clarity while optimally balancing information density and significantly minimizing mental effort. Furthermore, a human study with university students reveals that the visual summaries generated by KnowVis lead to substantially higher learning effectiveness and knowledge retention compared to baseline methods. Ultimately, this work demonstrates that intelligently transforming linear multimedia into structured visual narratives can significantly reduce cognitive barriers, paving the way for more accessible and engaging self-directed learning experiences.

## 2 KnowVis Framework

In this section, we formulate the task of generating visual summaries from video lectures and introduce our proposed KnowVis framework.

### 2.1 Problem Formulation

The primary objective of KnowVis is to generate visual summaries that facilitate learners' comprehension of complex key concepts within a video lecture. We formulate this task as a two-stage multimodal summarization problem tailored for educational contexts: $\mathcal{V} \rightarrow \mathcal{K} \rightarrow \mathcal{I}$.

**Stage 1: Video Lecture to Knowledge Units ($\mathcal{V} \rightarrow \mathcal{K}$).** Given an input video lecture $\mathcal{V}$, this stage constructs a set of discrete knowledge units $\mathcal{K}$ that encapsulate the essential multimodal information surrounding key concepts. We first extract

a comprehensive concept map $\mathcal{G}$ from $\mathcal{V}$ to capture the lecture's global knowledge structure. Within this structure, we identify key concepts by isolating structurally important concepts ($\mathcal{C}_{imp}$) and cognitively challenging concepts ($\mathcal{C}_{chg}$). We then construct a set of knowledge units $\mathcal{K}$, modeling each unit $K \in \mathcal{K}$ as the following tuple:

$$K = (c, G, T, S)$$

where $c \in (\mathcal{C}_{imp} \cup \mathcal{C}_{chg})$ represents the foundational key concept (e.g., "Lubrication"). $G \subseteq \mathcal{G}$ is a localized subgraph encoding the concept $c$ and its structurally cohesive relationships within $\mathcal{G}$. $T$ and $S$ denote the specific transcript segments and slide frames, respectively, temporally retrieved from $\mathcal{V}$ to semantically align with $c$.

**Stage 2: Knowledge Unit to Visual Summary ($\mathcal{K} \rightarrow \mathcal{I}$).** Given a structured knowledge unit $K \in \mathcal{K}$, this stage synthesizes the visual summary $I \in \mathcal{I}$. Instead of directly prompting an image generator with raw text, we translate the unit $K$ into a narrative visualization storyboard $B$. This storyboard maps abstract relationships to concrete visual metaphors and spatial layouts, which subsequently guides a text-to-image generative model $g$ to produce the final summary $I = g(B)$.

### 2.2 Framework Overview

As illustrated in Figure 1, KnowVis consists of three core modules. First, the *Concept Extraction module* (stage 1) processes multimodal video inputs to extract a fine-grained concept map $\mathcal{G}$. Next, the *Knowledge Units Construction module* (stage 1) evaluates this map to identify important and challenging concepts ($\mathcal{C}_{imp} \cup \mathcal{C}_{chg}$) and constructs localized knowledge units $\mathcal{K}$ by aggregating relevant subgraphs, transcripts, and slide frames. Finally, the *Visual Summarization Generation module* (stage 2) transforms these structured units into narrative visual summaries $\mathcal{I}$.

### 2.3 Concept Extraction

To build a comprehensive concept map $\mathcal{G}$, we first transcribe the lecture audio and employ a scene transition detection algorithm to isolate unique slide frames. To prevent long-context degradation, we segment video transcripts into discrete chunks aligned with these slide transitions. For each chunk, we employ a Multimodal Large Language Model (MLLM), i.e., Gemini-3-Flash (Google DeepMind, 2025), taking both the transcript text and the corresponding slide frames as input to extract localized concepts and their relations.

A persistent challenge in automated extraction is entity duplication, where the MLLM extracts semantically identical concepts using different phrasing. Inspired by (Wang et al., 2025b), we instruct the MLLM to perform entity resolution, aligning identical concepts both within and across temporal chunks. Furthermore, to mitigate MLLM hallucination, we apply a strict post-processing filter that prunes any concepts or relations not explicitly grounded in the source video lecture. Finally, isolated concept clusters lacking connections to the primary topic graph are removed to maintain narrative conciseness. Detailed extraction prompts are provided in Appendix A.1.

### 2.4 Knowledge Units Construction

Building upon $\mathcal{G}$, this module identifies critical pedagogical anchors and wraps them into cohesive knowledge units $\mathcal{K}$. Guided by Multimedia Learning Theory (Mayer, 2009) and the Threshold Concepts framework (Meek et al., 2023), we categorize key concepts into two dimensions:

- Important Concepts. Drawing from signaling principles (Mayer, 2009), we identify important concepts based on three criteria: 1) *Signaling*: presence of explicit visual/verbal emphasis; 2) *Temporal Allocation*: the duration of the instructor's focus; and 3) *Structural Connectivity*: the concept's centrality within the global map $\mathcal{G}$.
- Challenging Concepts. Informed by Threshold Concepts Theory (Meek et al., 2023), we define challenging concepts as those representing "troublesome knowledge". These are characterized by counterintuitive logic, highly technical jargon, or multi-step tacit reasoning that frequently disrupts novice comprehension.

We input the full transcripts, slides, and $\mathcal{G}$ into the MLLM, prompting it to score each concept across these dimensions. We select the top 10% of scoring nodes to form our final sets of important ($\mathcal{C}_{imp}$) and challenging ($\mathcal{C}_{chg}$) concepts (see Appendix A.2 for more implementation details).

**Subgraphs and Information Retrieval.** To construct the final knowledge unit $K$, we use each selected key concept $c \in (\mathcal{C}_{imp} \cup \mathcal{C}_{chg})$ as a seed node. We apply a label-spreading mechanism over $\mathcal{G}$ to cluster closely related neighboring concepts into a localized subgraph $G$. Because knowledge is inherently interconnected, concepts may belong

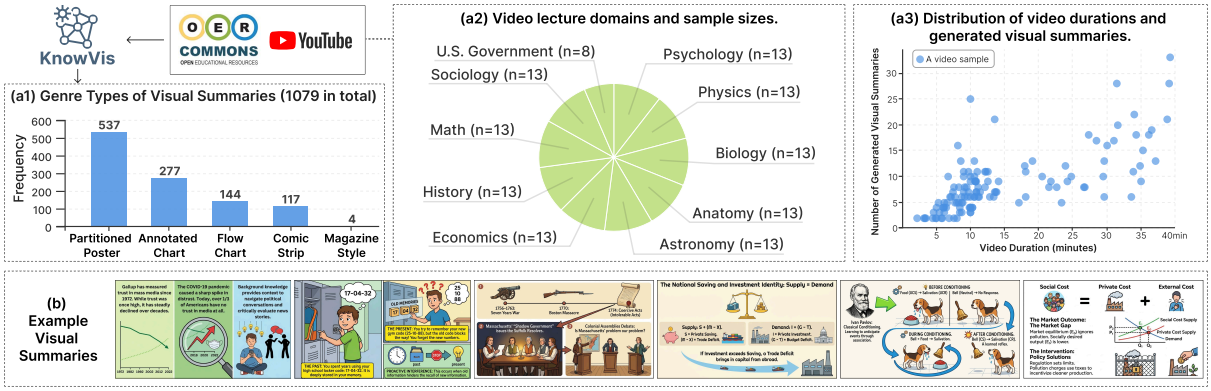


Figure 2: **Statistics and samples of our Dataset.** (a1–a3) Statistics of 1,079 visual summaries generated from 125 video lectures. (b) Samples illustrating diverse knowledge concepts (from left to right: "Trust in media", "Proactive interference", "Coercive acts", "Investment identity", "Ivan Pavlov", "Social cost") with pedagogically grounded visual narratives.

to multiple subgraphs if their association probabilities to multiple seeds exceed a threshold of 0.8 (hyperparameter details in Appendix A.3). Finally, we input the full transcripts, slides and $G$ into the MLLM to retrieve semantically aligned transcript segments $T$ and slide frames $S$, ensuring contextual completeness for the knowledge unit.

### 2.5 Visual Summarization Generation

Guided by the principles of narrative visualization (Segel and Heer, 2010), this module transforms the abstract data within $K$ into an intuitive visual summary $I$. First, we prompt Gemini-3-Flash to analyze content in $K$ to select the most pedagogically suitable visualization genre (e.g., a Flow Chart for procedural knowledge, an Annotated Chart for data, or a Comic Strip for abstract metaphors). The MLLM then generates a structured storyboard $B$, mapping the concepts in $G$ to specific visual layouts and descriptive captions.

This storyboard $B$ is passed to Gemini-3.1-Flash-Image (Google DeepMind, 2026) to synthesize the raw image. Because generative vision models occasionally produce typographical errors or hallucinated artifacts, we introduce an automated verification step. The generated image $I$ is fed back into Gemini-3-Flash alongside the original unit $K$; if semantic or visual inconsistencies are detected, the model refines the storyboard prompt $B$ and triggers a second generation pass to ensure high-fidelity educational accuracy (see Appendix A.4).

## 3 Dataset

To validate our framework, we required a domain-diverse, multimodal dataset. We curated a comprehensive collection of video lectures sourced from open-access platforms, primarily OER Commons[2] and YouTube[3]. To ensure content validity and pedagogical quality, we selected video lectures aligned with the curriculum of OpenStax[4] , a widely recognized open educational resource. To support future research, the collected videos and resulting data are openly licensed (see Appendix F).

As illustrated in Figure 2, the final dataset comprises 125 video lectures spanning 10 distinct academic disciplines. From these multimodal inputs, we utilized KnowVis to generate 1,079 pedagogically structured visual summaries. Each summary is paired with its source transcripts, slide frames, and concept graphs, providing a rich multimodal resource for future educational generation research.

## 4 Automatic Evaluation

Having constructed the dataset, we conduct a large-scale automated evaluation to quantitatively assess the pedagogical quality of the generated visual summaries. A case study showing qualitative comparisons is provided in Section 7.

### 4.1 Experimental Setup

**Metrics.** To quantitatively evaluate the generated visual summaries at scale, we adopt an LLM-as-a-judge approach, leveraging GPT-5.4 (OpenAI, 2026) as the automated evaluator (See detailed cross-model evaluation in Appendix C.4). Each visual summary is scored on a discrete scale from 1 to 5 across four pedagogical metrics. For the first two metrics, a higher score is strictly better:

[2] https://oercommons.org/
[3] https://www.youtube.com/
[4] https://openstax.org/

- **Accuracy:** Measures how faithfully the visual summary aligns with the factual information presented in the source video lecture.
- **Clarity:** Assesses how easily a novice student can interpret the visual summary without confusion or ambiguity (Wang et al., 2025a).

For the latter two metrics, the scoring follows a bell-curve preference where 3 is the balance. Scores of 1 (sparse/simple) or 5 (overwhelming/-complex) indicate suboptimal pedagogical design:

- **Information Density:** Measures the volume of information presented. Following Multimedia Learning Theory (Mayer, 2009), a visual summary should convey critical information (avoiding a score of 1) while omitting redundant details that cause clutter (avoiding a score of 5).
- **Mental Effort:** Assesses the cognitive load required to process the visual. According to Cognitive Load Theory (Sweller et al., 2011), the summary should foster active processing (avoiding 1) while minimizing extraneous cognitive load (avoiding 5).

The detailed prompt used for this LLM-based evaluation is provided in Appendix B.1.

**Baselines.** To evaluate the efficacy of KnowVis, we compare its performance against four baseline generation pipelines. The baseline pipelines represent different combinations of multimodal input and reasoning: **V2I** generates summaries directly from the raw video lecture file; **T2I** relies solely on the textual video transcripts; **TS2I** utilizes both transcripts and extracted slide frames; and **TS2I + CoT** incorporates an explicit Chain-of-Thought reasoning step prior to generation based on the transcripts and slides. To ensure a fair comparison, all baselines are strictly prompted to generate visual summaries for the exact same key concepts identified by our framework, which are clearer and more accurate than those extracted by baseline methods (see Appendix C.1).

**Model Constraints.** Because KnowVis utilizes Gemini-3.1-Flash-Image, we comprehensively evaluate this model across all four of the aforementioned pipelines. To establish a broader benchmark, we also compare our results against two other state-of-the-art models: Flux-2-Pro (Labs, 2025) and Qwen-Image-2 (Zhao et al., 2026). However, these models introduce specific technical constraints. Directly inputting video files (V2I) caused Out-Of-Memory (OOM) errors for 29 videos in our dataset, and both Flux and Qwen impose strict limitations on the number of image inputs they can process. Consequently, we evaluate Flux and Qwen exclusively using the text-only T2I pipeline. Furthermore, due to rigorous internal safety filters, Qwen-Image-2 refused to generate visuals for 51 sensitive concepts, while Flux-2-Pro failed for 442 concepts (primarily in Anatomy and History domains). We account for these failures in the results analysis by comparing methods using metric scores averaged over their successful generations. Appendix C.2 further provides intersection-based comparisons between KnowVis and these baselines.

### 4.2 Experiment Results

**Accuracy and Clarity.** We first evaluate the fundamental quality of the generated summaries. As shown in Table 1, KnowVis achieves the highest accuracy score (3.723), narrowly outperforming the top Gemini baselines. More notably, KnowVis yields a substantial improvement in clarity (4.470), leading all baseline methods by a wide margin. This significant gain highlights a core flaw in standard pipelines: baseline methods tend to dump all retrieved text and slide details into the prompt without prioritization, resulting in unfocused and chaotic images. By structuring knowledge units into a logical narrative flow before image synthesis, KnowVis aims to ensure that the core educational message remains focused and unambiguous.

**Information Density.** Next, we examine the volume of information presented, where a score of 3.0 represents the middle pedagogical density. While Qwen-T2I (2.994) and Flux-T2I (3.047) hover closest to this median, KnowVis scores slightly lower at 2.816. Rather than a deficiency, this conciseness is a deliberate and highly beneficial feature of our design. Because KnowVis distills lengthy transcripts into abstract concept graphs *before* generating the narrative, it inherently strips away verbal fluff. By prioritizing a coherent narrative structure over a high density of fine-grained details, the framework prevents the visual clutter that typically overwhelms novice learners.

**Mental Effort.** Finally, we assess the cognitive load imposed on learners. The visual summaries generated by KnowVis require significantly lower mental effort (2.366) to process than all other evaluated methods. Conversely, applying a CoT reasoning step to the TS2I pipeline generates the most mentally demanding summaries (3.201), as the

Table 1: **Quantitative evaluation of automated visual summary generation methods.** We compare direct and Chain-of-Thought (CoT) prompting across Flux, Qwen, and Gemini using various input modalities against our proposed KnowVis. KnowVis consistently outperforms baseline methods, achieving the highest accuracy and clarity while optimally filtering information density and significantly minimizing learner mental effort.

| Method | **Accuracy** ↑ | **Clarity** ↑ | **Information Density** | **Mental Effort** |
|---|---|---|---|---|
| Flux-T2I | 3.080 | 3.443 | 3.047 | 3.055 |
| Qwen-T2I | 3.618 | 4.152 | 2.994 | 2.764 |
| Gemini-T2I | 3.716 | 4.144 | 3.442 | 3.014 |
| Gemini-V2I | 3.564 | 4.102 | 3.335 | 2.981 |
| Gemini-TS2I | 3.721 | 4.057 | 3.471 | 3.083 |
| Gemini-TS2I (CoT) | 3.682 | 4.005 | 3.570 | 3.201 |
| KnowVis | **3.723** | **4.470** | 2.816 | 2.366 |

model attempts to visualize overly complex logical chains. The success of KnowVis lies in its synthesis of the previous metrics: high accuracy and clarity, combined with concisely filtered information density, naturally minimize cognitive fatigue. By wrapping facts in a structured narrative design, KnowVis actively reduces extraneous cognitive load, allowing learners to comprehend complex concepts with far less mental strain.

## 5 Ablation Study

To understand the function of each module in our pipeline, we conducted an ablation study. By systematically removing key modules, we can isolate their specific impact on pedagogical quality.

### 5.1 Experiment Setup

To understand the contribution of each module in our pipeline, we conducted an ablation study using Gemini-3-Flash and Gemini-3.1-Flash-Image across three degraded settings:

1. **w/o Concept Extraction:** We remove the structured map extraction, directly prompting the MLLM to extract text-based concepts from the transcripts and slides.
2. **w/o Knowledge Units Construction:** We skip the important/challenging filtering and label-spreading. The MLLM directly extracts subgraphs, retrieves text and slides to build units.
3. **w/o Visual Summarization Generation:** We bypass the narrative storyboard design, directly prompting the model with knowledge units.

### 5.2 Experiment Results

This section examines the role of each module by comparing metric scores averaged over generated visual summaries in each ablation setting (Table 2). A supplementary comparison based on video-level averages is provided in Appendix C.3.

**The Role of Concept Extraction.** We first evaluate the necessity of the global concept map. As shown in Table 2, without it, accuracy artificially rises slightly (3.723 → 3.738), but Clarity drops (4.470 → 4.354). Furthermore, the pipeline generated only 724 visual summaries compared to the baseline 1,079. When forced to extract concepts directly without a map, the MLLM outputs lengthy, convoluted sentences instead of precise keywords. While this smaller pool of broad content mathematically reduces the chance for generation errors (thereby raising Accuracy), these lengthy text strings fail to act as precise visual anchors. Consequently, the lack of structural precision significantly degrades the Clarity of the final image.

**The Role of Knowledge Units Construction.** Next, we assess the pedagogical filtering module. Skipping construction of key concepts results in massive, highly fragmented visual summaries (n=1,401). Because these summaries represent simple, isolated facts rather than structural knowledge, they score highest in Clarity (4.481) and lowest in Mental Effort (2.349), but suffer a drop in Accuracy (3.723 → 3.717). While isolated facts are superficially easier to read, they lack true pedagogical value. The Knowledge Unit module is critical because it forces the framework to anchor visual generation around threshold concepts, ensuring the summaries guide learners through actual, structured knowledge rather than disjointed trivia.

**The Role of Narrative Visual Generation.** Finally, we isolate the impact of narrative design by removing the Visual Summarization module. This removal causes a severe degradation in Clarity (4.470 → 4.068) while drastically increasing

Table 2: **Quantitative ablation study results.** We evaluate the impact of removing key structural modules from the KnowVis pipeline. The results demonstrate that each component plays a necessary role to balance high accuracy and clarity while maintaining optimal pedagogical cognitive load.

| **Setting** | Num.[a] | Accuracy ↑ | Clarity ↑ | Information Density | Mental Effort |
|---|---|---|---|---|---|
| KnowVis | 1079 | 3.723 | 4.470 | 2.816 | 2.366 |
| w/o Concept Extraction | 724 | 3.738 | 4.354 | 2.746 | 2.365 |
| w/o Knowledge Units Construction | 1401 | 3.717 | 4.481 | 2.857 | 2.349 |
| w/o Visual Summarization Generation | 1079 | 3.702 | 4.068 | 3.428 | 3.063 |

[a] Number of visual summaries generated.

Information Density (2.816 → 3.428) and Mental Effort (2.366 → 3.063). This demonstrates that directly feeding structured data to an image generator produces cluttered, overwhelming visuals. The intermediary step of designing a narrative storyboard is strictly essential to translate raw data into an intuitive, cognitively accessible format.

## 6 Human Evaluation

To evaluate the effectiveness of KnowVis in real-world learning environments, we conducted a user study with university students to capture authentic human interpretation of the generated visual summaries.

### 6.1 Experiment Setup

We recruited 10 university students ($P_1$–$P_{10}$) with diverse academic backgrounds, including Computer Science, Bioengineering, Architecture, Linguistics, Law, and Literature. The participants were randomly divided into five stable dyadic groups ($G_1$–$G_5$). We then randomly selected one video lecture from each of the 10 disciplines in our dataset, partitioned them into five pairs, and randomly assigned one pair to each group.

For each group, the two evaluators independently watched their assigned lectures and evaluated the corresponding visual summaries generated by KnowVis against two baseline configurations: T2I and TS2I using Gemini-3.1-Flash-Image. These baselines were selected as they demonstrated the strongest unimodal and multimodal performance, respectively, in our automated evaluation. The presentation order of the options was randomized, and evaluators were blind to the underlying methods to eliminate bias. Evaluators assessed the generation outputs across three core pedagogical dimensions:

- **Learning Effectiveness:** Measures how efficiently a student can comprehend the underlying target concept through the visual summary.
- **Knowledge Retention:** Assesses how reliably the core educational insights of the concept can be memorized and recalled via the visuals.
- **Knowledge Transfer:** Measures a student's willingness to share the visual summary with their peers as an illustrative learning aid.

Participants identified the best and worst visual summaries among the choices for each concept. To quantitatively analyze these subjective preferences, we mapped the categorical rankings to numerical scores: 5 for the best summary, 0 for the worst, and 3 for the intermediate runner-up. An example of the human evaluation interface is in Appendix E.

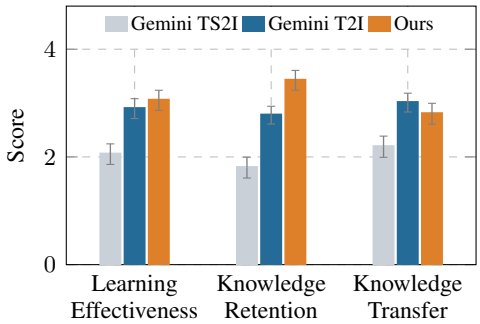


Figure 3: Average human-evaluated scores for generated visual summaries across Learning Effectiveness, Knowledge Retention, and Knowledge Transfer.

### 6.2 Experiment Results

As illustrated in Figure 3, KnowVis comfortably outperforms both baseline methods across the primary educational metrics. Specifically, for Learning Effectiveness, our framework achieves the highest average score (M=3.052, SE=0.186), followed by the T2I baseline (M=2.897, SE=0.185) and the TS2I configuration (M=2.052, SE=0.191). We observe a more pronounced advantage in Knowledge Retention, where KnowVis yields a clear lead (M=3.422, SE=0.184) over T2I (M=2.776, SE=0.165) and TS2I (M=1.802, SE=0.194). These

Table 3: Inter-Rater Consistency Evaluated by Weighted Cohen's Kappa Across Participants.

| Group | ID | Major | Learning Effectiveness ($\kappa$) | Knowledge Retention ($\kappa$) | Knowledge Transfer ($\kappa$) |
|---|---|---|---|---|---|
| $G_1$ | $P_1$ | Linguistic | 0.455 | 0.932 | 0.727 |
| | $P_2$ | Law | | | |
| $G_2$ | $P_3$ | Computer Science | $-0.096$ | $-0.096$ | 0.077 |
| | $P_4$ | Linguistic | | | |
| $G_3$ | $P_5$ | Literature | 0.250 | 0.156 | 0.156 |
| | $P_6$ | Public Health | | | |
| $G_4$ | $P_7$ | Bioengineering | 0.386 | 0.250 | 0.250 |
| | $P_8$ | Architecture | | | |
| $G_5$ | $P_9$ | Linguistic | 0.786 | 0.464 | $-0.286$ |
| | $P_{10}$ | Computer Science | | | |

findings corroborate our automated evaluation: by anchoring the visual generation around structured knowledge units and an explicit narrative flow, KnowVis provides sticky, low-effort cognitive landmarks that actively help students unpack and retain complex concepts. For the Knowledge Transfer metric, however, the text-only T2I baseline obtains a slightly higher average preference score (M=3.009, SE=0.175) than our framework (M=2.802, SE=0.195), while TS2I scores lowest (M=2.190, SE=0.197). To diagnose this variance, we analyze the inter-rater reliability below.

### 6.3 Consistency Analysis

We assessed inter-rater reliability using a linear-weighted Cohen's Kappa ($\kappa$). As shown in Table 3, Learning Effectiveness and Knowledge Retention maintain acceptable consistency across most dyads. Where divergence occurs, it is primarily driven by disparities in prior knowledge. For instance, in $G_2$ ($\kappa = -0.096$), an evaluator with high domain familiarity ($P_4$) preferred information-dense baselines to capture fine details, whereas a novice co-evaluator ($P_3$) heavily favored KnowVis's concise narratives to quickly grasp high-level concepts.

Inter-rater alignment drops further for Knowledge Transfer (e.g., $\kappa = -0.286$ in $G_5$), which post-study interviews attribute to competing social curation philosophies rather than pedagogical quality. While some students ($P_9$) prefer sharing KnowVis's lean summaries to spare peers from cognitive fatigue, others ($P_{10}$) share dense baseline visuals purely out of caution that peers might miss secondary details. This social dichotomy explains why text-heavy baselines marginally edge out KnowVis in raw transfer scores, despite lagging significantly in standalone educational utility.

## 7 Case Study

To qualitatively evaluate the visual summaries, we compare representative examples generated by different methods. As shown in Figure 4, KnowVis generates high-quality, pedagogically meaningful visual summaries that excel across multiple dimensions. First, compared with the baseline methods, KnowVis presents clearer and more intuitive textual and visual elements, enabling learners to identify essential information more efficiently. Next, it provides an appropriate amount of information without overwhelming details, avoiding the cognitive overload triggered by text-heavy diagrams. Furthermore, the content is highly accessible and easy to understand for novice learners. By integrating relatable real-world metaphors and using clear language, it bridges the gap between complex knowledge and intuitive understanding, maximizing its educational effectiveness for learners with diverse backgrounds.

## 8 Related Work

**Pedagogical Foundations.** The theoretical foundation of our work stems from the Cognitive Theory of Multimedia Learning and Cognitive Load Theory (Sweller et al., 2011; Sweller, 2020; Mayer and Moreno, 1998; Mayer, 2009). These theories demonstrate that processing linear instructional formats often overwhelms working memory, and that learning is optimized when verbal and visual representations are actively integrated. To systematically identify which concepts require this visual scaffolding, educational researchers rely on the Threshold Concepts framework (Meek et al., 2023) to pinpoint "troublesome knowledge" that acts as a bottleneck for novices. Our framework directly operationalizes these pedagogical theories to guide automated visual generation.

**Computational Approaches.** Existing computational efforts to mitigate learning barriers in video lectures generally fall into three categories. First, textual video summarization condenses multimedia content into text abstracts, either by generating global textual overviews (Gonzalez et al., 2023; Liu et al., 2025; Pennec et al., 2025), segment-anchored summaries (Lee et al., 2025; Lin et al., 2024), or synthesizing question-answer pairs that capture important topics (Ding et al., 2025; Ray et al., 2025). While this reduces overall length, it still outputs linear, text-heavy representations that

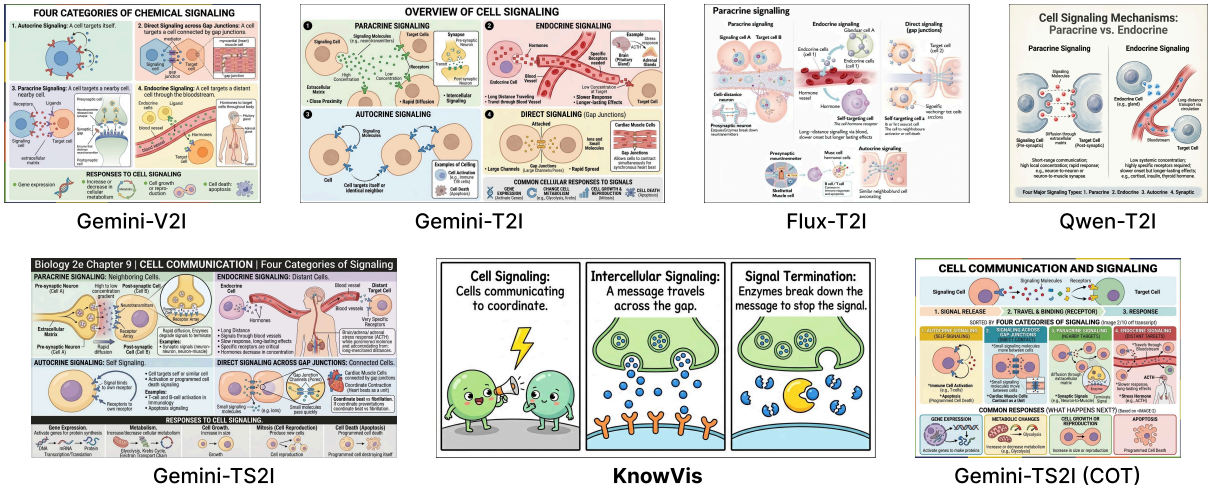


Figure 4: Qualitative comparison of visual summaries generated by different methods: direct and Chain-of-Thought (CoT) prompting across Flux, Qwen, and Gemini using various input modalities against our proposed KnowVis. KnowVis output optimizes learning effectiveness by delivering a concise, engaging visual narrative that is less cognitively demanding for learners.

fail to visually illustrate structural concept relationships. Second, video retrieval and indexing techniques enable learners to navigate to specific timestamps (Schwab et al., 2017; Zhou et al., 2025; Liu et al., 2018). However, these tools merely aid navigation without actively synthesizing the underlying knowledge. Finally, recent generative AI studies have explored domain-specific visual generation, such as math diagrams or medical flashcards (Wu et al., 2025; Wang et al., 2025a). While educationally effective, these approaches rely on heavily constrained, pre-defined templates that cannot generalize to the diverse, open-domain multimodal content of general lectures.

**Our Position.** Our proposed KnowVis addresses these limitations by bridging open-domain multimodal processing with pedagogically grounded visual synthesis. Unlike textual summarizers or navigational indices, KnowVis actively restructures linear video content into non-linear concept maps to expose underlying relationships. Furthermore, unlike domain-restricted visual generators, our framework utilizes MLLMs to dynamically design narrative visual summaries without relying on rigid templates, offering a scalable solution to reduce cognitive load in open-domain learning.

## 9 Conclusion

In this paper, we introduced KnowVis, a domain-independent framework that transforms lengthy, linear video lectures into pedagogically grounded visual summaries. By leveraging MLLMs to extract structural concept maps and identify challenging educational anchors, our pipeline generates intuitive narrative summarizations that explicitly map complex knowledge relationships. Extensive automated evaluations and human studies demonstrate that KnowVis significantly outperforms existing baselines, enhancing clarity and accuracy while effectively minimizing extraneous cognitive load. Ultimately, this work provides a scalable, multimodal solution to alleviate cognitive fatigue for novice learners, paving the way for more accessible and engaging self-directed education.

## Limitations

While our framework demonstrates strong potential for educational visual generation, we acknowledge several limitations that provide avenues for future research:

**Domain Variance (STEM vs. Non-STEM).** Although KnowVis is designed to be domain-independent, generating visual metaphors for highly abstract concepts in the Humanities or Social Sciences (e.g., Sociology, History) is inherently more subjective than visualizing concrete physical phenomena in STEM disciplines (e.g., Anatomy, Physics). In our current study, we observe that the visual generation models occasionally struggle to map abstract non-STEM concepts to intuitive graphical layouts. Future work should systematically investigate these cross-domain discrepancies and explore domain-adaptive prompting strategies.

**Dependence on Proprietary Models.** Our current pipeline relies on closed-source, proprietary models (Gemini-3-Flash and Gemini-3.1-Flash-Image) accessed via APIs. While these models represent the state-of-the-art in multimodal reasoning and generation, their closed nature introduces potential reproducibility challenges, as the models may be subject to silent updates. Furthermore, this limits the community's ability to fine-tune the underlying model weights on specific educational visual datasets. Future iterations should explore deploying open-weight MLLMs to ensure long-term reproducibility and adaptability.

**Scale of Human Evaluation.** Our user study provided rich qualitative insights regarding cognitive load, prior knowledge impacts, and social curation behaviors. However, the study was conducted with a relatively small sample size (N=10 university students) in a controlled setting. This limits the statistical generalizability of the human evaluation results. A critical next step is to conduct a large-scale, longitudinal deployment of KnowVis in real-world classroom environments to assess its impact on diverse learner populations over extended periods.

**Visual and Textual Generative Artifacts.** Despite implementing a closed-loop MLLM verification step to refine the storyboards, current text-to-image generative models still occasionally produce artifacts. These include typographic errors (e.g., misspelled text within an annotated chart) and minor spatial hallucinations. Fully eliminating these artifacts remains an open challenge in the broader generative AI field and currently requires human-in-the-loop validation for deployment in high-stakes educational settings.

**LLM-as-a-Judge Bias.** In our automatic evaluation, we utilized GPT-5.4 to conduct large-scale automated evaluations of our visual summaries. While recent literature validates this approach, LLM judges can exhibit inherent biases, such as preferring specific formatting styles or verbosity levels. Although our human evaluation corroborated the automated trends, the absolute automated scores should be interpreted as strong directional indicators rather than perfect objective truth.

## References


Mohamed Ally. 2004. Foundations of educational theory for online learning. *Theory and practice of online learning*, 2(1):15–44.

Mohamed Ally. 2008. *1. Foundations of Educational Theory for Online Learning*, pages 15–44. Athabasca University Press, Athabasca.

Anthropic. 2026. System card: Claude sonnet 4.6. Technical report, Anthropic.

Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. 2013. Studying learning in the worldwide classroom research into edx's first mooc. *Research & Practice in Assessment*, 8:13–25.

Chih-Ming Chen and Chung-Hsin Wu. 2015. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Comput. Educ.*, 80:108–121.

Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2025. Generating questions, answers, and distractors for videos: Exploring semantic uncertainty of object motions. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 7207–7220. Association for Computational Linguistics.

Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker, and Chris Callison-Burch. 2023. Automatically generated summaries of video lectures may enhance students' learning experience. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 382–393, Toronto, Canada. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash model card. Technical report, Google. Available at https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Google DeepMind. 2026. Gemini 3.1 flash image model card. Technical report, Google. Available at https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Image-Model-Card.pdf.

Raquel Gutiérrez-González, Alvaro Zamarron, and Ana Royuela. 2024. Video-based lecture engagement in a flipped classroom environment. *BMC Medical Education*, 24(1):1218.

Slava Kalyuga. 2005. Prior knowledge principle in multimedia learning. *The Cambridge handbook of multimedia learning*, pages 325–337.

Black Forest Labs. 2025. Flux.2: Frontier visual intelligence. Technical report, Black Forest Labs. Available at https://bfl.ai/models/flux-2.

Min Jung Lee, Dayoung Gong, and Minsu Cho. 2025. Video summarization with large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18981–18991. Computer Vision Foundation / IEEE.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Trans. Multim.*, 26:5548–5560.

Ching (Jean) Liu, Juho Kim, and Hao-Chuan Wang. 2018. Conceptscape: Collaborative concept mapping for video learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 387. ACM.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. What is that talk about? a video-to-text summarization dataset for scientific presentations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6187–6210, Vienna, Austria. Association for Computational Linguistics.

Richard E. Mayer. 2009. *Multimedia Learning*, 2nd edition. Cambridge University Press, USA.

Richard E Mayer and Roxana Moreno. 1998. A cognitive theory of multimedia learning: Implications for design principles. *Journal of educational psychology*, 91(2):358–368.

Sarah E. M. Meek, Hilary Neve, and Andy Wearn. 2023. *Threshold Concepts and Troublesome Knowledge*, pages 361–383. Springer Nature Singapore, Singapore.

OpenAI. 2026. Openai GPT-5 system card. *CoRR*, abs/2601.03267.

Galann Pennec, Zhengyuan Liu, Nicholas Asher, Philippe Muller, and Nancy F. Chen. 2025. Integrating video and text: A balanced approach to multimodal summary generation and evaluation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pages 2403–2426. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Sourjyadip Ray, Shubham Sharma, Somak Aditya, and Pawan Goyal. 2025. Eduvidqa: Generating and evaluating long-form answers to student questions based on lecture videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 34701–34727. Association for Computational Linguistics.

Daniel L Schacter and Karl K Szpunar. 2015. Enhancing attention and memory during video-recorded lectures. *Scholarship of Teaching and Learning in Psychology*, 1(1):60.

Michail Schwab, Hendrik Strobelt, James Tompkin, Colin Fredericks, Connor Huff, Dana Higgins, Anton Strezhnev, Mayya Komisarchik, Gary King, and Hanspeter Pfister. 2017. booc.io: An education system with hierarchical concept maps and dynamic non-linear learning plans. *IEEE Trans. Vis. Comput. Graph.*, 23(1):571–580.

Daniel T. Seaton, Yoav Bergner, Isaac Chuang, Piotr Mitros, and David E. Pritchard. 2014. Who does what in a massive open online course? *Commun. ACM*, 57(4):58–65.

Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148.

John Sweller. 2020. Cognitive load theory and educational technology. *Educational technology research and development*, 68(1):1–16.

John Sweller, Paul Ayres, and Slava Kalyuga. 2011. Cognitive load theory.

Junling Wang, Anna Rutkiewicz, April Yi Wang, and Mrinmaya Sachan. 2025a. Generating pedagogically meaningful visuals for math word problems: A new benchmark and analysis of text-to-image models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 11229–11257. Association for Computational Linguistics.

Tianshu Wang, Xiaoyang Chen, Hongyu Lin, Xuanang Chen, Xianpei Han, Le Sun, Hao Wang, and Zhenyu Zeng. 2025b. Match, compare, or select? an investigation of large language models for entity matching. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 96–109. Association for Computational Linguistics.

Anna Wong, Wayne Leahy, Nadine Marcus, and John Sweller. 2012. Cognitive load theory, the transient information effect and e-learning. *Learning and Instruction*, 22(6):449–457.

Qian Wu, Zheyao Gao, Longfei Gou, Yifan Hou, Ann Sin Nga Lau, and Qi Dou. 2025. Healthcards: Exploring text-to-image generation as visual aids for healthcare knowledge democratizing and education. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 27536–27558. Association for Computational Linguistics.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, Kun Yan, Liang Peng,

Lihan Jiang, Niantong Li, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, and 56 others. 2026. Qwen-image-2.0 technical report. *Preprint*, arXiv:2605.10730.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhiguang Zhou, Li Ye, Lihong Cai, Lei Wang, Yigang Wang, Yongheng Wang, Wei Chen, and Yong Wang. 2025. Conceptthread: Visualizing threaded concepts in MOOC videos. *IEEE Trans. Vis. Comput. Graph.*, 31(2):1354–1370.

# A Methodological Details of the KnowVis Framework

## A.1 Extracting Concept Map

Table 4 presents the step-by-step prompting methods for extracting concept maps from the transcript chunks and corresponding slide frames.

## A.2 Important and Challenging Concepts Identification

Figure 5 shows the evaluation criteria used to score the importance of concepts, and Table 6 shows the evaluation criteria used to identify the challenging concepts.

**Prompt for Scoring Concept Importance**

**Step 1-Analyze:** Please evaluate the importance of each concept based on the following four criteria:
**1. Multimedia Signaling:** Whether the concept is visually emphasized on slides (e.g., primary heading, large arrows, or highlighting).
**2. Lexicogrammatical Relevance Markers:** Whether verbal cues are used to emphasize a concept's high importance (e.g., verb patterns ("Must remember,"), noun phrases (e.g., "The most important," "the key thing"), and evaluative adjectives (e.g., "crucial," "fundamental")).
**3. Structural Connectivity:** Whether the concept is pivotal to the network's connectivity (i.e., if this concept were removed, would the logical flow or the integration of other sub-topics be significantly compromised?).
**4. Temporal Allocation:** Whether the instructor invests significant narrative effort to define, elaborate, or revisit it across multiple segments or slides.
Evaluate and score each concept based on the four defined criteria.

**Step 2-Rank:** Rank the candidate concepts from highest to lowest weighted importance.

Figure 5: Illustration of the prompt used to score the importance of the concepts in the concept map

**Prompt for Identifying Challenging Concepts.**

**Step 1-Analyze:** Please evaluate whether each candidate concept is challenging for student Alex (a university student with little prior knowledge of the subject, politically neutral and objective) based on the following eight criteria:
**1. Counterintuitive:** Whether the concept goes against students' current conceptual model of the topic.
**2. Inert:** Whether the concept can not fit easily into learners' current understanding or may conflict with their world view or beliefs, e.g., a student may not understand why anyone would smoke when they know that smoking causes disease, until the student encounters someone who experiences little power or agency.
**3. Alien:** Whether the student cannot connect the concept to the world around them, so then struggles to apply it meaningfully in practice or in new contexts.
**4. Ritual:** Whether the concept is primarily encoded as routine answers to common questions, making it difficult for students to grasp the underlying principles? For instance, students may be able to perform a respiratory examination exactly as taught, yet lack understanding of the underlying concepts and the rationale for each step.
**5. Troublesome language:** Whether the concept includes medical jargon which may be meaningless to students and render a topic incomprehensible.
**6. Conceptually difficult:** Whether the concept is complex in nature. For example, many medical concepts are extremely complex.
**7. Tacit:** Whether the concept is implicit within a discipline but may not be overtly explained during the video lecture.
**8. Nettlesome:** Whether the concept challenges cultural or individual beliefs and results in an intense emotional response.
Select the candidate concepts satisfying at least one of the eight criteria. Discard low-level details; retain only challenging concepts that are important for learning.
**Step 2-Rank:** Rank the selected challenging concepts from highest to lowest level of challenging based on the eight criteria.

Figure 6: Illustration of the prompt used to identify challenging concepts from the concept map.

## A.3 Knowledge Units Construction Algorithm

Algorithm 1 shows the detailed procedure for extracting subgraphs to construct knowledge units via label propagation.

Note that the LabelSpreading model is configured with a maximum of 1,000 iterations, a convergence tolerance of $10^{-3}$, and a clamping factor of $\alpha = 0.9$. Here, $\alpha = 0.9$ ensures that nodes depend more on the surrounding graph structure than on their original seed labels. This allows certain seed concepts to be absorbed by seeds, capturing the inherent hierarchical nature of knowledge, in which concepts can be subsumed or nested.

## A.4 Visual Summarization Generation

We present prompts for generating visual summaries from knowledge units in Table 5.

# B Experiment Details

## B.1 LLM-as-a-judge

Figure 7 presents the evaluation prompt and scoring rubrics we used for LLM-as-a-judge to assess the visual summaries.

Table 4: Prompts for Extracting Concept Map.

| **Step 1 - Concept Extraction** |
| --- |
| Focus on the given chunk and the corresponding slides. The task is to extract knowledge concepts from the provided sentences in the chunk. Knowledge concepts should be informative and meaningful entities with clear semantic significance, no more than three words. |
| **Step 2.1 - Relation Extraction** |
| Focus on concepts in CL, explore relationships between concepts based on the sentences in the given chunk, and the corresponding slides. Generate knowledge triples among these concepts in the format (head, relation, tail), where the 'head' and 'tail' are concepts in CL, and 'relation' represents the extracted relationship between them. |
| **Step 2.2 - Checking Missing Triples** |
| Focus on all the heads and tails in this JSON and the previous JSON; do you miss any relationships among them that are indicated or mentioned in and across all the given chunks and the corresponding slides, such as inclusion and other logical relationships? |
| **Step 2.3 - Checking Triples Content** |
| Check the triples based on the following two principles for constructing triples: Each triple should represent a complete proposition with a clear and direct relational description. The head and tail of the triple should be distinct knowledge concepts in CL, each concept should be an informative and meaningful entity with clear semantic significance, and no more than three words. If any triples do not meet these principles, refine them accordingly to ensure compliance based on the given chunk. |
| **Step 3.1 - Entity Matching within Chunks** |
| Check this JSON for triples where the heads and tails may have different descriptions but refer to the same real-world entity. Identify duplicated entities that are functionally identical due to minor formatting, punctuation, or spelling variations (e.g., "5 number summary" and "Five-number summary"). Only select entities that unambiguously represent the same object or label without conceptual nuance. Do not group technical synonyms (e.g., "Q1" and "25th percentile") where distinct terminology serves an educational purpose. |
| **Step 3.2 - Entity Matching across Chunks** |
| Refer to triples across all chunks. Identify duplicated entities that are functionally identical due to minor formatting, punctuation, or spelling variations (e.g., "5 number summary" and "Five-number summary"). Only select entities that unambiguously represent the same object or label without conceptual nuance. Do not group technical synonyms (e.g., "Q1" and "25th percentile") where distinct terminology serves an educational purpose. |

### B.2 Baseline Methods

We present the prompts used for the baseline methods to generate visual summaries in Table 6.

## C More Experimental Results and Analysis

### C.1 Concepts Evaluation

We use GPT-5.4 (OpenAI, 2026) to compare the concept list extracted by KnowVis and direct prompting for all videos in our dataset. As Table 7 shows, our module extracts clearer, more accurate concepts than the baselines, which demonstrates that using KnowVis-extracted key concepts as a controlled variable actually benefits the baselines by providing high-quality anchors. This confirms that our controlled evaluation design is reasonable for isolating the quality of visual summary generation from differences in concept selection.

### C.2 Supplementary Analysis of Automatic Evaluation Results

Table 1 reports the main comparison by averaging scores over all generated visual summaries for each method. Since some baseline models encountered failures or refusals (Gemini-V2I, Qwen-T2I, and Flux-T2I), we stratified the results in Table 1 into different subsets based on the intersection of concepts and recalculated the average scores for each subset. As shown in Table 8, our method consistently outperforms the baseline models across all baseline-specific valid subsets.

### C.3 Supplementary Analysis of Ablation Study Results

Section 5 reports the ablation results by averaging scores over all generated visual summaries in each setting. Since different ablation settings yield varying numbers of visual summaries, we supplement the results in Table 2 with a video-level comparison by averaging the scores within each video and then

Table 5: Prompts used by KnowVis to generate the visual summaries in Figure 1.

| **Step 1: Visual Storyboard Generation** |
|---|
| Generate a concise storyboard for a visual summary explaining "lubrication" for a novice learner. Focus on illustrating the knowledge triples. Take the transcript and slide images as the factual basis.<br>Please follow the steps below:<br>1. GENRE & STRATEGY: Choose the most appropriate genre (Magazine Style, Annotated Chart, Partitioned Poster, Flow Chart, or Comic Strip) to best illustrate the core concepts.<br>2. OUTLINE: Generate a concise storyboard outline by synthesizing the provided knowledge triples into their most essential visual narrative arc. Be faithful to the source material. Use the minimum number of panels to capture the core logic without unnecessary details.<br>3. PANEL BREAKDOWN: For each panel, provide:<br>• Panel ID: (e.g., Panel-1)<br>• Panel Triple To Be Focused: (Note which triples this panel focuses on)<br>• Panel Text To Be Rendered: (Provide the EXACT, concise text that should appear in the panel)<br>• Panel Visual Description: (Detailed description of the visual elements to be illustrated) |
| **Step 2: Visual Summarization Generation** |
| Generate a visual educational summary for the concept 'lubrication' based on the given visual storyboard outline. Follow the outline strictly. DO NOT add any extra text or numbers outside the storyboard. Create a visually appealing design. Avoid serious or boring vibes; keep it vivid, approachable, and clear. Enrich the illustration with informative visual metaphors and expressive icons, without adding extra text. Render text in a legible format using a friendly font. Don't restrict to a rigid grid layout. Arrange the panels' layout flexibly to optimize narrative flow and visual balance. |
| **Step 3: Post-Verification** |
| Analyze the provided visual summary for the following errors:<br>1. Repetitions: Duplicated text, redundant panels, or ghosting elements.<br>2. Logical inconsistency across elements (count, equations, cause-and-effect).<br>If errors are found: Output ONLY a revised prompt using the Location-Based Structuring method that fixes these issues by removing duplications and correcting logical inconsistencies. If no errors: Output ONLY the word 'PASS'. |

**Algorithm 1 Extract Knowledge Unit Subgraphs from a Concept Map**

**Input:** Concept Map $G = (V, E)$, Seed Node Set $C_{\text{seed}} = C_{\text{imp}} \cup C_{\text{chg}} \subseteq V$.
**Output:** Set of Knowledge Units $K$

Execute Label Propagation on Concept Map $G$:
$P \leftarrow \text{LabelSpreading}(G, C_{\text{seed}})$, where $P \in \mathbb{R}^{|V| \times |C_{\text{seed}}|}$
**for all** $v \in V$ **do**
  $M_v \leftarrow \max_{s \in C_{\text{seed}}} P_{v,s}$
**end for**
**for all** $s \in C_{\text{seed}}$ **do**
  $V_s \leftarrow \{v \in V \mid P_{v,s} \geq 0.8 \times M_v\} \cup \{s\}$
  $V_s \leftarrow \text{ConnectedComponent}(G[V_s])$ // Filter out isolated nodes
  $K \leftarrow K \cup \{G[V_s]\}$
**end for**
$K \leftarrow \{S \in K \mid \forall S' \in K \setminus \{S\}, E(S) \not\subseteq E(S')\}$ // Deduplication
**return** $K$

aggregating them across the dataset. The results in Table 9 are consistent with Table 2: each component is necessary to balance high accuracy and clarity while maintaining optimal cognitive load.

### C.4 Cross-Model Validation of Automatic Evaluation

We conduct a supplementary evaluation with Claude-Sonnet-4.6 (Anthropic, 2026) on a random 10% subset of our dataset, covering 111 visual summaries from 17 videos. The results evaluated by Claude-Sonnet-4.6, shown in Table 10, indicate that KnowVis achieves the highest accuracy and clarity while maintaining lower mental effort, showing a trend similar to the main results in Table 1. To further examine cross-model consistency, we calculate the agreement between GPT-5.4 and Claude-Sonnet-4.6 scores across all evaluated outputs ($N = 111 \times 7$). As shown in Table 11, the two evaluators exhibit high within-one-score agreement across all metrics, further supporting the robustness of our automatic evaluation.

## D Computational Experiments

We implemented our proposed framework using Gemini-3-Flash and Gemini-3.1-Flash-Image. To ensure reproducibility, the temperature for Gemini-3-Flash was set to 0. For Gemini-3.1-Flash-Image, we set the temperature to 1.0 to promote generation creativity and stylistic diversity, and configured the thinking level to High to leverage its maxi-

Figure 7: Prompts for LLM-as-a-judge of the generated visual summaries

| | |
|---|---|
| **Accuracy**<br>How accurately the visual illustration content factually align with the source video transcript and slides.<br>**1:** Very inaccurate. Contains major factual contradictions that fundamentally conflict the source.<br>**2:** Inaccurate. Contains multiple factual errors that are likely to cause significant misunderstanding.<br>**3:** Generally accurate. Largely correct, but contains at least one noticeable factual error.<br>**4:** Accurate. Highly correct, with only negligible errors in unimportant details.<br>**5:** Perfectly accurate. Entirely correct; every visual and textual detail is precise without error. | **Clarity**<br>**Definition:** How easily students can interpret the visual illustration without confusion or ambiguity.<br>**1:** Very unclear. Interpretation is nearly impossible.<br>**2:** Unclear. Significant content ambiguity or poor visual organization makes interpretation difficult.<br>**3:** Generally Clear. Content is generally understandable, though minor ambiguities or visual clutter may slow comprehension.<br>**4:** Clear. Content is easy to understand, with only negligible ambiguities that do not impact comprehension.<br>**5:** Very Clear. Content is entirely intuitive. Optimal visuals and zero ambiguity ensure immediate and accurate comprehension. |
| **Information Density**<br>The amount of information presented in the visual illustration.<br>**1:** Contains no important information.<br>**2:** Contains minimal information and lacks the most important source details.<br>**3:** Contains sufficient information and captures the essential takeaways.<br>**4:** Contains highly detailed information.<br>**5:** Contains overwhelmingly dense and excessively detailed information. | **Mental Effort**<br>How much mental effort is required for a learner to process the visual illustration. **1:** Fails to trigger active cognitive processing.<br>**2:** Minimal mental effort; the content is effortless to process.<br>**3:** Moderate mental effort; the content demands a manageable amount of cognitive load to process.<br>**4:** High mental effort; the content demands a high level of cognitive load to process.<br>**5:** Mental overload; the content demands excessive cognitive load, leading to mental exhaustion. |

mum reasoning capabilities. The only exception is the post-verification regeneration step within our framework. In this step, the thinking level is set to Minimal to ensure the newly generated visuals strictly and honestly align with the refined script. For automatic evaluation, we compared our approach against six state-of-the-art baselines utilizing Gemini-3.1-Flash-Image, Flux-2-pro, and Qwen-Image-2. We also conducted an ablation study using both Gemini-3-Flash and Gemini-3.1-Flash-Image. We used the same parameters (i.e., temperature, thinking level) for our framework and baseline methods.

The total cost incurred for API queries across all development, ablation, and evaluation experiments was approximately $1000. In practice, processing a 40-minute video via the KnowVis pipeline takes about 30 minutes and costs 3 USD for API queries. The actual latency and cost can fluctuate depending on the information density and complexity of the lecture content.

## E Details of Human Evaluation

An example of the human evaluation interface presented to the participants is provided in Figure 8. Participants were recruited from a local university via word of mouth. The study was conducted remotely via Zoom, with each individual session lasting approximately 45 minutes. Each participant received a $10 compensation after the study.

All human evaluation protocols were approved under the ethics approval ID: `HU-STA-00001449`. Participants were informed of the research purpose and study procedures, and they provided informed consent for the publication of their evaluation results. No personally identifiable information was collected. Furthermore, all experimental materials were derived from open educational resources using open-access models. There is no risk of introducing offensive content or private data leaks.

Table 6: Baseline method prompt examples for generating the visual summaries in Figure 1.

| **Prompts for Baseline V2I** |
|---|
| Based on the provided video, create a visual educational illustration for the concept 'Lubrication.' |
| **Prompts for Baseline T2I** |
| Based on the provided transcript, create a visual educational illustration for the concept 'Lubrication.' |
| **Prompts for Baseline TS2I** |
| Based on the provided transcript and video slides, create a visual educational illustration for the concept 'Lubrication.' |
| **Prompts for Baseline TS2I (COT)** |
| Step one: analyze the provided transcript and video slides to summarize the core knowledge of the concept 'Lubrication.'<br>Step two: Based on the analysis, create a visual educational illustration for the concept 'Lubrication.' |

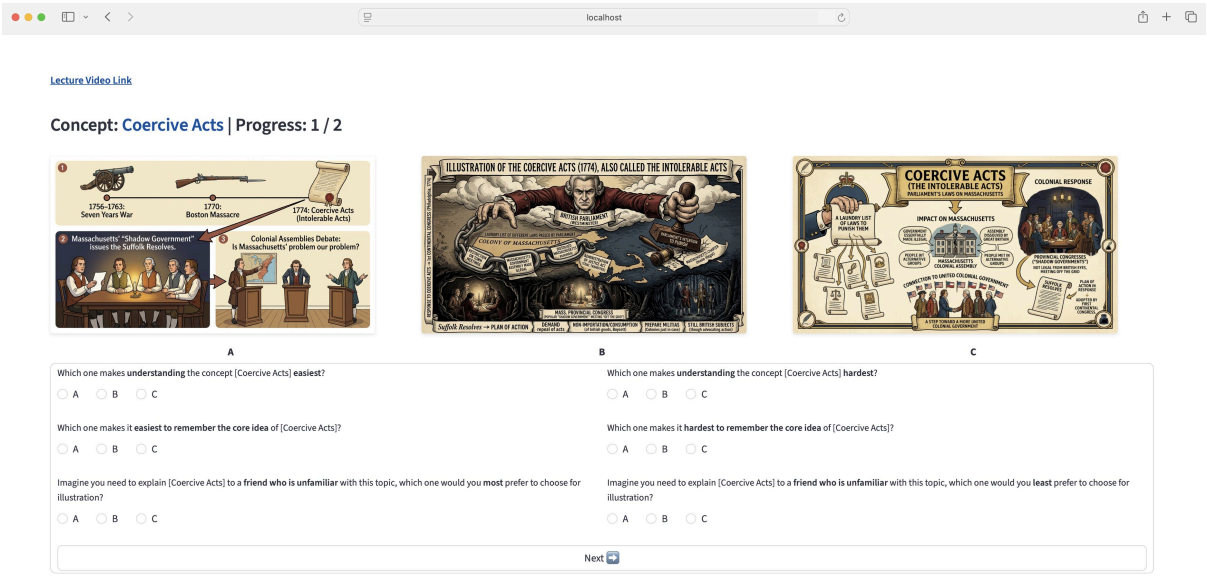


Figure 8: The user interface designed for human evaluation for comparison between our generated visual summary (A) and the baselines, including Gemini-TS2V (B) and Gemini-T2V (C). To ensure an unbiased assessment, the order of the options is randomly shuffled, and participants remain blinded to the underlying methods.

Table 7: Quantitative evaluation of our concept extraction method. KnowVis outperforms the baseline by extracting clearer and more accurate concepts.

| Method | **Accuracy** ↑ | **Clarity** ↑ |
|---|---|---|
| Direct LLM Prompting | 4.072 | 3.472 |
| **KnowVis** | **4.144** | **3.888** |

## F Use or Create Scientific Artifacts

We curated our dataset by sourcing educational video lectures directly from open-access channels. These video lectures span diverse disciplines, including Physics, Astronomy, Anatomy, Psychology, Math, Sociology, Biology, Economics, U.S. History, and American Government. The curated course playlists are publicly available via the following open-access channels: Introduction to Physical Science, Introduction to Astronomy, Anatomy and Physiology I and II, Introduction to Psychology 1e, Introductory Statistics, Introduction to Sociology 3e, Biology2E, Economics 3rd Edition, U.S. History Volume I: Chapters 1-16, and American Government. We used open educational resources and open-access models to generate visual summaries for education. No new content was created by the researchers. There is no risk of introducing offensive content or personally identifying information.

The video lectures selected for our dataset are shared under permissive terms, including the Creative Commons Attribution license (CC BY, reuse allowed) and the Creative Commons Attribution — NonCommercial — ShareAlike 4.0 International License (CC BY-NC-SA 4.0), or are governed by the Standard YouTube License. Data access and

Table 8: Supplementary subset comparison of automatic evaluation results across visual summary generation methods that encounter failures or refusals for some input concepts. Each subset is defined by the intersection of concepts for which visual summaries are successfully generated, and scores are averaged within each subset.

**KnowVis vs. Gemini-V2I**

| Method | Accuracy ↑ | Clarity ↑ | Information Density | Mental Effort |
|---|---|---|---|---|
| Gemini-V2I | 3.564 | 4.102 | 3.335 | 2.981 |
| KnowVis | **3.732** | **4.477** | 2.852 | 2.404 |

**KnowVis vs. Qwen-T2I**

| Method | Accuracy ↑ | Clarity ↑ | Information Density | Mental Effort |
|---|---|---|---|---|
| Qwen-T2I | 3.618 | 4.152 | 2.994 | 2.764 |
| KnowVis | **3.720** | **4.462** | 2.813 | 2.372 |

**KnowVis vs. Flux-T2I**

| Method | Accuracy ↑ | Clarity ↑ | Information Density | Mental Effort |
|---|---|---|---|---|
| Flux-T2I | 3.078 | 3.443 | 3.047 | 3.055 |
| KnowVis | **3.732** | **4.480** | 2.846 | 2.403 |

Table 9: Supplementary Analysis of Ablation Study Results at the video level comparison. Metric scores are averaged within each video to evaluate the individual contribution of each framework component to the generated visual summaries.

| **Setting** | Accuracy ↑ | Clarity ↑ | Information Density | Mental Effort |
|---|---|---|---|---|
| KnowVis | **3.738** | **4.487** | 2.872 | 2.414 |
| w/o Concept Extraction | 3.687 | 4.289 | 2.784 | 2.449 |
| w/o Knowledge Units Construction | 3.685 | 4.474 | 2.877 | 2.354 |
| w/o Visual Summarization Generation | 3.671 | 4.066 | 3.460 | 3.063 |

Table 10: Quantitative evaluation of visual summary generation methods using Claude-Sonnet-4.6 on a random 10% subset of the dataset.

| Method | **Accuracy ↑** | **Clarity ↑** | **Information Density** | **Mental Effort** |
|---|---|---|---|---|
| Flux-T2I | 2.910 | 3.190 | 2.937 | 2.856 |
| Qwen-T2I | 2.991 | 3.523 | 2.730 | 2.478 |
| Gemini-T2I | 3.559 | 3.793 | 3.306 | 3.081 |
| Gemini-V2I | 3.396 | 3.784 | 3.234 | 2.910 |
| Gemini-TS2I | 3.532 | 3.604 | 3.441 | 3.234 |
| Gemini-TS2I (COT) | 3.577 | 3.505 | 3.703 | 3.441 |
| KnowVis | **3.658** | **4.027** | 2.702 | 2.387 |

Table 11: Consistency between GPT-5.4 and Claude-Sonnet-4.6 scores on the evaluated subset ($N = 111 \times 7$).

| Metric | Exact Agreement | Within-1 Agreement | Weighted Cohen's $\kappa$ |
|---|---|---|---|
| Accuracy | 0.537 | 0.972 | 0.415 |
| Clarity | 0.430 | 0.950 | 0.288 |
| Information Density | 0.671 | 0.960 | 0.450 |
| Mental Effort | 0.622 | 0.945 | 0.383 |

content analysis were conducted strictly through YouTube's public Application Programming Interfaces (APIs) for non-commercial evaluation, ensuring full compliance with the platform's Terms of Service and academic reuse guidelines. Our use of these scientific artifacts is consistent with their intended pedagogical and public educational purposes. To support future study, our curated dataset will be released to the research community under the CC BY-NC-SA 4.0 license.

## G AI Assistants In Research Or Writing

We used AI assistants such as Gemini exclusively for language editing and grammar refinement.