KCVR: Knowledge-Centric Video Reconstruction for Structured Pedagogical Summarization via Dynamic Graph Planning

Jingjiang Liu[1*], **Jia Zhu**[1*†], **Hanghui Guo**[2], **Weijie Shi**[3], **Yue Cui**[4], **Xiaokang Jin**[1], **Yilin Wang**[1], **Qingyu Niu**[1], **Jiawei Shen**[1], **Guoqing Ma**[1], **Yidan Liang**[1], **Shimin Di**[2], **Jiajie Xu**[5†]

[1]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University
[2]School of Computer Science and Engineering, Southeast University
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[4]Alibaba Group, [5]School of Computer Science and Technology, Soochow University

Abstract

Existing video summarization methods mainly compress content for gist browsing, but they often break the prerequisite logic in instructional videos and induce logical inversions (e.g., conclusions before premises). We formalize this problem as Structure-Pedagogical Reconstruction (**SPR**). SPR raises two challenges: (1) **Structure Hallucination**, where retrieved knowledge is topologically valid but not evidence-grounded by the blackboard; and (2) **Logical Inversion**, where soft prompt-level graph injection fails to enforce prerequisite order during decoding. To address these challenges, we propose **K**nowledge-**C**entric **V**ideo **R**econstruction **(KCVR)**, a Plan-then-Generate neuro-symbolic framework that decouples epistemic planning from content generation. KCVR prunes a Dual-Layer Epistemic Graph into a minimal video-supported plan, then realizes the plan with visually anchored attention and topology-constrained decoding. We additionally release **EduStruct**, a 10-discipline benchmark for SPR and structure-centric evaluation. Experiments show that KCVR outperforms strong end-to-end baselines on Knowledge Progression Consistency and Learning Objective Coverage. Our code and data are available at https://github.com/mark1001-ljj/video_sum.

1 Introduction

Effective summarization of instructional videos is pivotal for scalable knowledge dissemination and retention (Ackermans et al., 2025; Xu et al., 2025). Existing video summarization models are predominantly designed for "gist browsing", extracting salient clips from unstructured streams like vlogs to provide a surface-level overview (Li et al., 2023; Fu et al., 2025). However, instructional videos demand *Pedagogical Reconstruction*: recovering the instructor's implicit logical chain (e.g., definition → theorem → proof) rather than merely compressing content. Consider a math lecture deriving the Arithmetic Sequence Formula: the transcript is often filled with deictic references (e.g., "from this, we get that"), while the crucial derivation steps are conveyed only through the spatial layout of blackboard visuals (Munasinghe et al., 2023). In such contexts, the challenge is not just selecting keyframes, but reconstructing a coherent, prerequisite-consistent knowledge path in which visual evidence grounds verbal reasoning (Tang et al., 2025). Formally, we term this task **S**tructure-**P**edagogical **R**econstruction (**SPR**), defined as generating a single structured summary with an explicit plan (i.e., an ordered concept trajectory) that

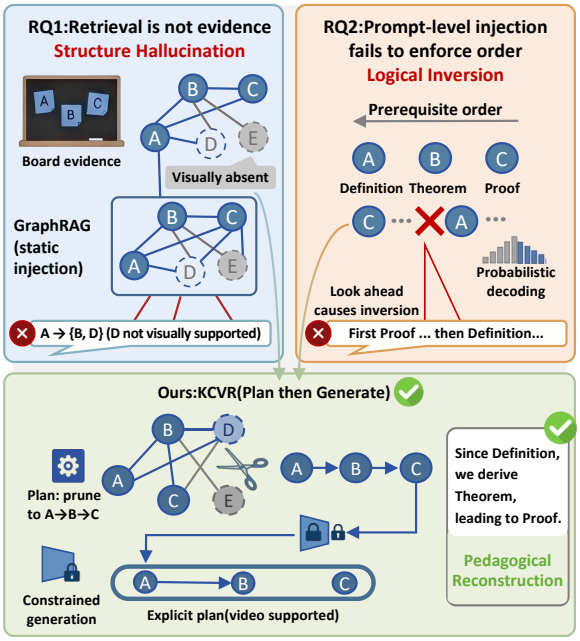

Figure 1: **Failure modes in SPR and our solution**. Static GraphRAG induces structure hallucination and probabilistic decoding causes logical inversion; KCVR resolves both via Plan then Generate.

*These authors contributed equally.
†Corresponding author.
e-mail: 15579996895@zjnu.edu.cn, jiazhu@zjnu.edu.cn, xujj@suda.edu.cn

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9133–9160
July 2-7, 2026

respects prerequisite order, grounded in blackboard visual evidence.

While Multimodal Large Language Models (MLLMs) (He et al., 2024) have demonstrated impressive capabilities in general video comprehension, applying them to SPR exposes a misalignment between probabilistic decoding and pedagogical topology(Guo et al., 2025; Zhu et al., 2025). Specifically, classroom blackboards often encode strict prerequisite-ordered derivations, yet Video LLMs still yield **Logical Inversion** errors (Ren et al., 2024; Zhang et al., 2024), hallucinating conclusions before premises. Unlike open-ended video summarization, SPR demands strict prerequisite-ordered reasoning grounded in visual evidence (Netland et al., 2025). Since such reasoning fundamentally relies on explicit dependency representations that exceed the implicit capabilities of probabilistic models, Knowledge Graphs (KGs) emerge as the natural candidate to provide the necessary structural prior.

However, effectively integrating KGs into MLLMs (Liu et al., 2025) for SPR presents a distinct challenge: it is not merely about retrieving *more* knowledge (Liu et al., 2024; Gao et al., 2024), but about enforcing an explicit structural prior that is both evidence-grounded and prerequisite-consistent. Current KG-augmented MLLM pipelines typically adopt a passive "retrieve-then-append" paradigm, treating the KG simply as an external context buffer rather than a logical constraint. They directly serialize local KG neighborhoods into the prompt, often ordered by semantic relevance rather than pedagogical topology. This semantic-first approach fundamentally misaligns with the causal nature of SPR: without an active mechanism to align visual evidence with topological rules, MLLMs remain prone to two systematic failures illustrated in Figure 1: *Structure Hallucination* from evidence-free retrieval and *Logical Inversion* under probabilistic decoding.

RQ1: In SPR, retrieval is *not* evidence: a KG neighborhood may be topologically valid yet visually absent from the blackboard. When GraphRAG-style retrieval is driven by semantic proximity, the injected subgraph becomes an over-complete implicit plan and induces **Structure Hallucination**. The key question is how to dynamically prune the graph into a minimal, video-supported prerequisite path before generation.

RQ2: Even with a correct plan, **Logical Inversion** can persist because prompt-level graph injection remains a soft condition during decoding; the model can still "look ahead" and verbalize future nodes before their prerequisites. How can we enforce step-wise generation to follow the planned prerequisite order, suppressing future concepts without sacrificing fluency?

To address these two failure modes, we propose **K**nowledge-**C**entric **V**ideo **R**econstruction **(KCVR)**, a neuro-symbolic pipeline that separates evidence-aware planning from topology-aware realization under a "Plan-then-Generate" principle. First, to align reconstruction with the natural rhythm of instruction, we parse the video into canonical pedagogical phases (e.g., Introduction, Exposition). Then, to mitigate Structure Hallucination, we introduce the **Subgraph Generator Planner (SG-Planner)**. Unlike static retrieval, SG-Planner acts as a dynamic filter that employs Contextual Pruning to actively strip away branches inconsistent with the current transcript and blackboard frames, synthesizing a precise lesson plan. Finally, to prevent Logical Inversion, we instantiate a *topology-constrained decoding* mechanism with two coupled modules: (1) **K**nowledge-**G**uided **V**isual **A**ttention **(KGVA)**, which grounds the current concept to specific blackboard regions; and (2) **A**daptive **C**onstrained **P**edagogical Decoding (**ACP**), which modulates the token distribution to turn the recovered plan from a soft prompt into an explicit topological prior. This substantially improves prerequisite-consistent progression while preserving linguistic fluency, effectively curbing the probabilistic freedom that leads to logical errors. The main contributions are listed as follows:

- We formalize **SPR**, a new task that shifts the goal of educational video summarization from content compression to the recovery of explicit, prerequisite-consistent knowledge paths.
- We propose **KCVR**, the first Plan-then-Generate framework for SPR. By introducing the SG-Planner for dynamic graph pruning and ACP for constrained decoding, KCVR effectively mitigates logical inversion and structure hallucination errors inherent in probabilistic MLLMs.
- We release **EduStruct**, a large-scale multimodal benchmark covering 10 disciplines (460+ videos), accompanied by Dual-Layer Epistemic Graphs that decouple curriculum flow from concept logic to support structure-centric research.

- We establish a structure-centric evaluation protocol via Knowledge Progression Consistency (KPC) and Learning Objective Coverage (LOC). Experiments validate KCVR's superiority, achieving statistically significant improvements over strong baselines in pedagogical rigor.

2 Related Work

2.1 Video Summarization Paradigms.

Multimodal summarization has evolved from visual-saliency approaches like DSNet (Zhu et al., 2021) and SUM-GAN (Apostolidis et al., 2021) to cross-modal alignment methods. Recent works like CLIP-It (Narasimhan et al., 2021) and MF2Summ (Wang and Zhang, 2025; Hu et al., 2023) leverage joint embeddings for semantic grounding, while query-focused models like VideoXum (Lin et al., 2024) introduce textual guidance. However, these paradigms predominantly treat videos as flat temporal streams, which excel at spotting visual highlights but cannot model the rigid prerequisite chains essential for pedagogical reconstruction (Hua et al., 2025).

2.2 Video LLMs for Long Context.

Video-LLMs have advanced long-form comprehension. Models like Video-LLaMA2 (Zhang et al., 2023) and mPLUG-Owl (Ye et al., 2024) enable multi-modal perception via instruction tuning. To handle rich temporal contexts, MovieChat (Song et al., 2024) introduces sparse memory mechanisms, while Chat-UniVi (Jin et al., 2024) unifies visual tokens for efficiency. Others, such as Video-ChatGPT (Maaz et al., 2024) and InternVL (Chen et al., 2024; Bagheri et al., 2024; Kar et al., 2025), enhance detailed understanding through large-scale alignment. Without explicit curriculum constraints, however, these probabilistic models risk generating locally plausible but pedagogically misordered explanations in classroom videos.

2.3 Structure-Aware Generation.

Integrating symbolic constraints into neural generation is gaining traction. NeuroLogic Decoding (Lu et al., 2021, 2022) and its variants pioneered the use of logical predicates to guide beam search. Recently, GraphRAG (Han et al., 2025) and GCR (Luo et al., 2025) demonstrated how retrieving structured subgraphs can ground LLM reasoning. Other works like JointGT (Ke et al., 2021; Tu et al., 2024; Huang et al., 2025) explore structure-aware text generation. While effective for unimodal text, these methods do not address multimodal video streams where constraints must reflect pedagogical precedence across visual-verbal evidence. In contrast, our KCVR enforces curriculum logic while preserving linguistic fluidity in educational video reconstruction.

3 Methodology

3.1 Framework Overview & Structural Priors

Figure 2 summarizes KCVR as a latent-plan factorization for SPR. We first segment a lecture into pedagogical phases via P^3, yielding $\{(T_i, I_i)\}_{i=1}^N$. Given each phase segment, Stage 1 maps $(T_i, I_i, \mathcal{G})$ to an ordered, evidence-supported plan $\mathcal{P}_i = (V_{\text{sub}}^{(i)}, E_{\text{sub}}^{(i)}, \pi^{(i)})$. Stage 2 realizes $\mathcal{P}_i$ into a micro-summary s_i by coupling KGVA, which conditions visual attention on the active concept in $\pi^{(i)}$, and ACP, which constrains decoding by suppressing strictly future concepts in $\pi^{(i)}$. Stage 3 fuses $\{s_i\}_{i=1}^N$ under a Global Knowledge Sketch constructed from the union of $\{\mathcal{P}_i\}_{i=1}^N$ to produce structured lecture notes.

Definition 1: Dual-Layer Epistemic Graph. We formalize domain knowledge as a dual-layer epistemic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Appendix B) that decouples *narrative scope* from *epistemic precedence*. Nodes are partitioned into curriculum units $\mathcal{V}_{tp}$ and knowledge concepts $\mathcal{V}_{kc}$, corresponding to TeachingPoint and KnowledgeConcept in our schema. The **Curriculum Layer** $\mathcal{V}_{tp}$ anchors scope through structural relations (e.g., PART_OF) and teaching-sequence relations (e.g., NEXT).

The **Concept Layer** $\mathcal{V}_{kc}$ encodes directed dependencies: BASED_ON captures definitional or derivational precedence, while sparse PREREQ_OF edges mark high-penalty prerequisite barriers. Cross-layer alignment uses HAS_CONCEPT edges from TeachingPoints to KnowledgeConcepts, enabling scope-constrained candidate filtering followed by dependency-consistent ordering in planning.

Definition 2: Pedagogical Phase Parsing. We adopt four canonical pedagogical phases, *Introduction*, *Exposition*, *Interaction*, and *Conclusion*. We implement a **Pedagogical Phase Parser (P^3)** as a transcript-only sequence labeling model that assigns a phase label to each utterance and converts the labels into temporal boundaries, yielding phase segments $\{(T_i, I_i)\}_{i=1}^N$ for downstream planning. At inference time, P^3 takes only the raw transcript

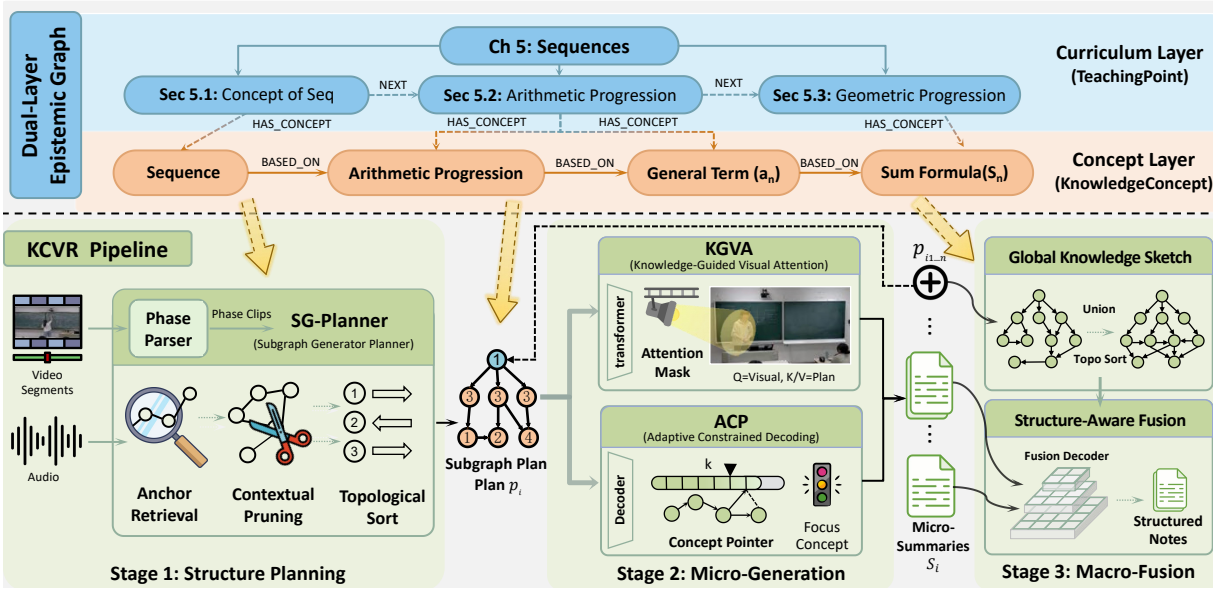

Figure 2: Detailed framework of **K**nowledge-**C**entric **V**ideo **R**econstruction **(KCVR)**. Specifically, Stage 1: SG-Planner prunes the Dual-Layer Epistemic Graph into an ordered, video-supported plan π. Stage 2: KGVA grounds concepts to blackboard regions and ACP performs topology-constrained decoding to generate phase summaries. Stage 3: GKS fuses micro-summaries into structured lecture notes.

as input and uses no video-specific annotations. To avoid evaluation leakage, P^3 is trained and selected without using any video from the summarization test split; implementation details and segmentation statistics are provided in Appendix A.2.

3.2 Stage 1: Pedagogical Structure Planning

Interface. Stage 1 maps a phase segment (T, I) and graph $\mathcal{G}$ to an ordered, evidence-supported plan $\mathcal{P} = (V_{\text{sub}}, E_{\text{sub}}, \pi)$ for downstream generation.

SG-Planner addresses RQ1 by turning retrieval into an explicit plan that is minimal and video-supported. SG-Planner mitigates structure hallucination by pruning visually absent yet topologically plausible branches before generation.

Inputs and evidence. Given a phase segment with transcript T and sampled frames I, we construct multimodal evidence E and compute relevance scores $s_{tp}(v)$ and $s_{kc}(u)$ via dense retrieval over node textual fields (name+definition). We select the curriculum anchor $v^*_{tp} = \arg\max_{v \in \mathcal{V}_{tp}} s_{tp}(v)$; mapping rules are in Appendix A.4.

Pedagogical anchoring and candidate set. The anchor v^*_{tp} defines a curriculum scope $\Phi(v^*_{tp}) \subseteq \mathcal{V}_{tp}$ via PART_OF and NEXT relations on the TeachingPoint layer. We construct a candidate set by combining evidence retrieval and local expansion:

$$\mathcal{U}_{\text{cand}} = \textsc{Expand}\big(\textsc{TopK}(s_{kc}), \mathcal{G}, h\big) \quad (1)$$

where expansion follows BASED_ON edges. We define curriculum prior $\mathcal{C}(u) = \mathbb{I}[u \text{ is linked to } \Phi(v^*_{tp}) \text{ via HAS\_CONCEPT}]$.

curriculum-aware pruning. Retrieval relevance alone does not guarantee visual support, so we enforce minimality under a curriculum scope prior. To filter out concepts that are logically relevant but unsupported by the current video evidence, we formulate pruning as a budgeted selection problem:

$$V^*_{\text{sub}} = \arg \max_{S \subseteq \mathcal{U}_{\text{cand}}, |S| \le B} \sum_{u \in S} (s_{kc}(u) + \lambda \cdot \mathcal{C}(u)) \quad (2)$$

Here, λ is a hyperparameter regulating the trade-off between evidence relevance and curriculum alignment. We empirically set $\lambda = 0.5$ based on validation performance; a detailed sensitivity analysis demonstrating robustness across $\lambda \in [0.1, 1.0]$ is provided in Figure 8 (Appendix C). This explicit scoring objective trades off local evidence against global curriculum scope, offering a deterministic alternative to opaque neural selection.

Minimum-violation ordering. To prevent plan-level inversions induced by relevance-ranked serialization, we derive precedence constraints from BASED_ON edges within V^*_{sub} and linearize them

Algorithm 1: SG-Planner: evidence-aware subgraph planning

Require: Epistemic graph $\mathcal{G}$, evidence E, budget B, hop h
Ensure: Plan $\mathcal{P} = (V_{\text{sub}}, E_{\text{sub}}, \pi)$
1: $v^*_{tp} \leftarrow \arg\max_{v \in \mathcal{V}_{tp}} s_{tp}(v)$
2: $U_{\text{cand}} \leftarrow \textsc{Expand}(\textsc{TopK}(s_{kc}), \mathcal{G}, h)$
3: $V_{\text{sub}} \leftarrow \arg\max_{S \subseteq U_{\text{cand}}, |S| \le B} \sum_{u \in S} \big(s_{kc}(u) + \lambda \cdot \mathcal{C}(u)\big)$
4: $E_{\text{sub}} \leftarrow \{(u, u') \in \texttt{BASED\_ON} : u, u' \in V_{\text{sub}}\}$
5: $\pi \leftarrow \textsc{MinViolationOrder}(V_{\text{sub}}, E_{\text{sub}})$
6: **return** $\mathcal{P}$

with minimal violations. We derive precedence constraints from BASED_ON edges within the selected concept set U^*_{sub}. Let $E^{\rightarrow}_{kc}$ denote the set of such edges. We seek an ordering π that minimizes the weighted violation cost:

$$\pi^* = \arg\min_{\pi} \sum_{(u,u') \in E^{\rightarrow}_{kc}} \mathbb{I}[\pi(u) > \pi(u')] \cdot w_{u,u'} \tag{3}$$

where $w_{u,u'}$ assigns penalties to reversing logical dependencies. Since this ordering problem is NP-hard (equivalent to the Feedback Arc Set problem on general graphs), we approximate the solution via a greedy local search initialized by s_{kc}-based sorting. This is efficient for small pedagogical subgraphs ($|U_{\text{sub}}| \le B$) and empirically recovers prerequisite-consistent plans;implementation details are in Appendix C.

3.3 Stage 2: Knowledge-Guided Micro-Generation

Given the plan $\mathcal{P} = (V_{\text{sub}}, E_{\text{sub}}, \pi)$, Stage 2 addresses **RQ2** by turning a symbolic trajectory into step-wise, evidence-grounded text generation. The key challenge is that plan injection alone is a soft condition: a Video-LLM may still attend to irrelevant regions or verbalize future concepts ahead of their prerequisites. We thus couple plan-guided perception (KGVA) with plan-constrained decoding (ACP), aligning visual evidence with the current concept while suppressing premature look-ahead.

KGVA: Plan-guided visual grounding. KGVA injects the semantic intent of π into the visual encoder to resolve deictic and visually implicit teaching cues. We represent the plan as concept embeddings $C_\pi = [\mathbf{e}_{u_1}, \ldots, \mathbf{e}_{u_m}]$, and insert a gated cross-attention layer where visual tokens X_v serve as Queries ($Q = X_v W_Q$) and C_π serves as Keys/-Values ($K = C_\pi W_K, V = C_\pi W_V$). The visual features are updated as:

$$X'_v = X_v + \alpha \cdot \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V, \tag{4}$$

where α is a zero-initialized gate for stable training. This plan-guided attention suppresses visual distractions and amplifies blackboard regions consistent with the active concept; architecture and visualizations are in Fig. 10 (Appendix D).

ACP: Adaptive Constrained Pedagogical Decoding. While KGVA grounds perception, decoding must follow the prerequisite order in π to avoid *Epistemic Leakage*. ACP maintains a concept pointer k and modulates logits $\mathbf{z}_t$ to promote aliases of the current concept while inhibiting aliases of strictly future concepts:

$$\mathbf{z}'_t = \mathbf{z}_t + \beta \cdot \mathbf{1}_{\mathcal{A}(\pi_k)} - \gamma \cdot \mathbf{1}_{\mathcal{F}(\pi_{>k})}. \tag{5}$$

Here, $\mathcal{A}(\pi_k)$ is the alias set of surface forms of the current concept π_k, and $\mathcal{F}(\pi_{>k}) = \bigcup_{j>k} \mathcal{A}(\pi_j)$ collects aliases of strictly future concepts. The pointer advances when the cumulative probability on $\mathcal{A}(\pi_k)$ exceeds a threshold or when explicit discourse markers (e.g., "Next") are generated, enforcing concept completion before transition.

3.4 Stage 3: Global Knowledge Sketch

Stage 2 produces phase-level micro-summaries, but pedagogical coherence is a global property: prerequisite chains and learning goals can span multiple phases. A naive concatenation of $\{s_1, \ldots, s_N\}$ yields *Structural Drift*, where key teaching points are omitted or re-ordered across long-range dependencies. To stabilize the macro reconstruction, we introduce the **G**lobal **K**nowledge **S**ketch **(GKS)**.

We build a global structural scaffold $\mathcal{K}_{\text{global}}$ by topologically sorting the union of planner-verified phase plans $\{\mathcal{P}^{(i)}\}_{i=1}^N$, where $\mathcal{P}^{(i)} = (V^{(i)}_{\text{sub}}, E^{(i)}_{\text{sub}}, \pi^{(i)})$. Concretely, we construct $\mathcal{K}_{\text{global}}$ from the union graph with nodes $\bigcup_i V^{(i)}_{\text{sub}}$ and dependency edges $\bigcup_i E^{(i)}_{\text{sub}}$. Because $\mathcal{K}_{\text{global}}$ is derived from planner-verified nodes and dependencies, it provides an explicit global constraint that is independent of the generator's local fluency.

Finally, a structure-aware generator fuses $\{s_1, \ldots, s_N\}$ conditioned on $\mathcal{K}_{\text{global}}$. The generator performs narrative threading to improve transitions while explicitly instructed to cover the nodes in $\mathcal{K}_{\text{global}}$, reducing omission and reordering errors in long-range aggregation. Micro-level evidence for phase-wise gains is in Appendix A.5.

4 Experiments

4.1 Experimental Setups

Dataset: The EduStruct Benchmark. To evaluate structure-aware reasoning in high-density pedagogical contexts, we introduce EduStruct, a large-scale benchmark curated from 463 expert-level classroom videos (≈95h). Distinct from generic video datasets dominated by visual events, EduStruct features **Dual-Layer Epistemic Annotation**, hierarchically parsing lessons into 1,852 phase segments (463 videos × 4 phases) grounded to 9k knowledge nodes (see Figure 3 for distribution).

To rigorously test generalization, we implement a Leave-Three-Subjects-Out **(L3SO)** protocol:

- **Source Domains (Train/Val)**: 7 subjects (e.g., Math, Physics) with explicit structural priors to learn meta-pedagogical schemas.
- **Target Domains (Test)**: 3 held-out subjects (**History**, **English**, **InfoTech**) to assess zero-shot structural transfer.

Crucially, the InfoTech subset serves as an *Adversarial Structural Testbed*: consisting of live coding sessions; we evaluate it under a **KG-withheld** inference setting to test structure induction from learned pedagogical priors. It forces the model to synthesize structure solely from latent pedagogical patterns learned from source domains, strictly penalizing "structure hallucination."

KG availability and evaluation-only signals. Under L3SO, *History* and *English* are evaluated with their subject graphs available at inference, while *InfoTech* is evaluated in a strict **KG-withheld** setting where no external knowledge graph is provided to the model at test time. For all splits, KPC/LOC are computed offline using the held-out annotation contract (learning objectives and knowledge points), which are **never** exposed as model inputs; this evaluation contract remains available even when the subject KG is withheld at inference (e.g., InfoTech), and thus does not imply any test-time knowledge input to the model; details on leakage control are in Appendix A.3.

Implementation Details. We employ InternVL3-8B as the primary backbone due to its high-resolution visual encoding, with VideoLLaMA2-7B used to verify model-agnosticism. Crucially, our main setting does not use explicit OCR, relying solely on visual grounding (OCR variants are in Appendix E.4). Detailed hyperparameters (e.g., pruning budget $B = 20$) are listed in Table 17 (Appendix E.4). Statistical significance is assessed via paired Wilcoxon signed-rank tests; the substantial effect sizes (Cohen's $d > 0.8$ for KPC gains) confirm robustness against multiple hypothesis testing corrections.

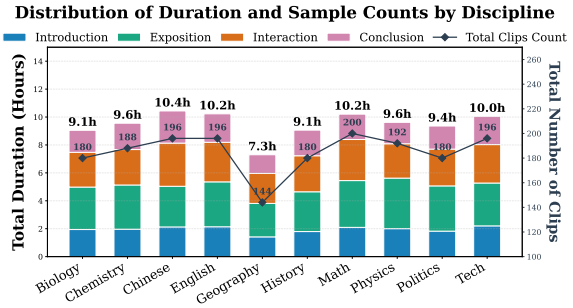

Figure 3: **EduStruct dataset statistics.** Total duration per discipline is shown as stacked bars over four pedagogical phases (Introduction, Exposition, Interaction, Conclusion), while the line plot indicates the total number of clips for each discipline.

Baselines. We compare KCVR against four categories of baselines to validate each component: (i) **Transcript-only LLM:** Qwen2.5-Instruct$_{\text{transcript}}$ assesses the text-only reference without visual or graph guidance. (ii) **End-to-End Video-LLMs:** We evaluate *InternVL3* and *VideoLLaMA2* in both Zero-Shot and SFT settings. (iii) **Long-Context Approach:** *Long-Context Stuffing* concatenates the full transcript with the same sampled frames as our main setting (same backbone and frame budget), and truncates the combined prompt to the model limit (up to 32k tokens) to test whether raw context length can supersede explicit planning. (iv) **Retrieval-Augmented Generation (RAG):** We compare *Text-RAG* (retrieved definitions) and *GraphRAG* (retrieved subgraphs). To ensure fair comparison, *GraphRAG* uses a DPR retriever capped at Top-$K = 20$ (matching our planner's budget) over the same retrieval corpus and node textual fields (name+definition) as KCVR, and injects triples ordered by retrieval scores, lacking the topological linearization of KCVR.

4.2 Metric Validation via Human Correlation

Before reporting main results, we rigorously validate that our proposed graph-based metrics (KPC, LOC) align with expert judgment. We conducted a human evaluation on 50 randomly sampled videos (stratified across subjects and visual information density), where three expert educators rated sum-

Method	Param	Input			Generation				Pedagogical		
		T	V	KG	R-1	R-2	R-L	BS	C-F1	KPC	LOC
GPT-4o	-	✓	✓	-	35.66	12.55	30.81	82.11	63.55	52.40	50.68
Qwen2.5-Instruct$_{\text{transcript}}$	7B	✓	-	-	31.98	9.82	19.93	80.27	60.97	44.33	48.35
InternVL3 (Zero-shot)	8B	✓	✓	-	34.60	10.63	26.34	79.62	60.45	46.69	46.77
VideoLLaMA2 (Zero-shot)	7B	✓	✓	-	33.12	10.20	25.40	79.10	59.28	45.17	45.36
InternVL3 (SFT)	8B	✓	✓	-	41.20	14.77	34.51	81.54	65.43	56.52	54.89
VideoLLaMA2 (SFT)	7B	✓	✓	-	40.59	13.45	32.87	80.87	64.80	55.28	53.41
Long-Context Stuffing	8B	✓	✓	-	41.83	15.10	33.25	81.96	66.25	57.45	55.56
MapReduce (no KG)	8B	✓	✓	-	40.86	14.98	32.10	80.07	67.30	54.78	53.69
Text-RAG (Top-K)	8B	✓	-	✓	41.96	15.65	33.66	81.35	67.59	59.82	57.88
GraphRAG (Standard)	8B	✓	✓	✓	42.23	15.80	34.53	82.14	68.69	60.55	58.52
VideoLLaMA2 + KCVR*	7B	✓	✓	✓	**42.72**	**16.04**	**34.22**	**82.16**	**69.88**	**66.15**	**62.00**
InternVL3 + KCVR (Ours)*	8B	✓	✓	✓	**43.68**	**16.68**	**36.62**	**83.56**	**70.83**	**66.79**	**63.11**

Table 1: **In-domain results (N=48).** T=transcript, V=frames, KG=retrieval corpus; BS=BERTScore$\times$100. Text-RAG: Top-K concept definitions; GraphRAG: Top-K subgraphs (score-ordered, no topology-aware linearization). Macro-avg over 3 seeds; $^{*}p < 0.001$ vs. the backbone-matched SFT baseline (paired Wilcoxon, video-level).

maries on *Pedagogical Coherence* and *Goal Fulfillment* (1-5 Likert scale); protocol details are in Appendix E.2. As shown in Table 2, KPC exhibits a strong correlation ($\rho = 0.72$) with human coherence ratings, showing higher correlation than generic metrics like ROUGE-L ($\rho = 0.41$) and BERTScore ($\rho = 0.48$). This supports KPC as a reliable automated proxy for pedagogical coherence, motivating its use as a primary structure metric in our evaluation.

Auto Metric	Human Dimension	Spearman ρ
ROUGE-L	Pedagogical Coherence	0.41
BERTScore	Pedagogical Coherence	0.48
KPC (Ours)	**Pedagogical Coherence**	**0.72****
LOC (Ours)	**Goal Fulfillment**	**0.68****

Table 2: **Human Correlation Study.** KPC shows higher correlation with pedagogical coherence than ROUGE-L and BERTScore; LOC correlates with goal fulfillment ($^{**}p < 0.01$ for testing $\rho \neq 0$).

4.3 Main Results: In-Domain Evaluation

Table 1 summarizes in-domain results (N=48).

Planning matters beyond long context. Long-context stuffing improves ROUGE-L (33.25) but still underperforms on pedagogical structure (KPC 57.45). Our full model reaches 66.79 KPC and 63.11 LOC ($p < 0.001$), suggesting that prerequisite consistency requires explicit structure rather than more context.

KCVR transfers across backbones. Applying KCVR to VideoLLaMA2 raises KPC from 55.28 to 66.15 (+10.87) and LOC from 53.41 to 62.00 (+8.59), surpassing even InternVL3 (SFT) on KPC. This indicates that the gains stem mainly from the framework components, not the backbone.

Dynamic planning beats static injection. Static retrieval baselines (Text-RAG/GraphRAG) improve coverage but lack topological ordering. The Full Model outperforms them by **+6.97** and **+6.24 KPC** respectively, isolating the critical value of the planner's trajectory optimization.

Visual input correlates with higher LOC. The best vision-agnostic baseline (Text-RAG) reaches 57.88 LOC. Structured KGVA boosts this to **63.11 (+5.23)**, confirming that guided attention effectively unlocks blackboard evidence that text-only methods miss.

4.4 Out-of-Domain Generalization (L3SO)

We evaluate zero-shot structural transfer on 3 held-out domains (Table 3), revealing a clear generalization hierarchy. (1) KG-Supported Transfer (History/English): KPC drops moderately (6.5%–8.3%) relative to in-domain but consistently outperforms SFT, indicating robust adaptation to unseen topologies. (2) KG-Absent Transfer (InfoTech): In this adversarial setting without graph access, SG-Planner relies solely on internalized pedagogical priors (e.g., Code follows Concept). Remarkably, the Full Model achieves 59.67 KPC (+10.7 vs. SFT), retaining $\sim$89% of in-domain performance. This confirms that KCVR learns transferable pedagogical meta-schemas rather than merely memorizing triples.

Subject	SFT KPC	SFT LOC	Full KPC	Full LOC	ΔKPC
In-Domain	56.52	54.89	**66.79**	**63.11**	-
History (w/ KG)	49.21	47.65	**61.23**	**58.04**	-8.3%
English (w/ KG)	50.37	48.72	**62.45**	**59.28**	-6.5%
InfoTech (no KG)	48.95	47.12	**59.67****	**56.92**	**-10.7%**
OOD Avg	49.51	47.83	**61.12**	**58.08**	**-8.5%**

Table 3: **L3SO Zero-Shot Transfer (N=143 OOD videos).** Train: 7 ID subjects (320 videos). *InfoTech (no KG)*: no graph at inference. $^{\star\star}p < 0.01$ vs SFT (Wilcoxon). Full Model retains about 89% ID KPC on InfoTech (no KG).

Variant	Pln	KGVA	ACP	R1	C-F1	KPC
InternVL3 (SFT)	-	-	-	41.20	65.43	56.52
+ Text-RAG	-	-	-	41.96	67.59	59.82
+ Planner	✓	-	-	42.60	68.40	62.10
+ KGVA only	✓	✓	-	42.96	69.05	63.22
+ ACP only	✓	-	✓	42.85	69.20	64.43
+ KGVA+ACP	✓	✓	✓	43.30	69.80	65.60
Full (w/ GKS)	✓	✓	✓	**43.68**	**70.83**	**66.79**

Table 4: **Module Contributions** (N=48 validation videos). Largest single-step KPC gain: Planner (+2.28). With the planner fixed, ACP-only yields the larger single-module KPC gain, while KGVA+ACP performs best. $p < 0.01$ vs. prior (Wilcoxon test).

Under KG-withheld inference, Stage 1 keeps the same planning interface $P_i = (V_i^{sub}, E_i^{sub}, \pi_i)$ but degenerates into a plan-induction mode that does not access any external subject graph. Concretely, V_i^{sub} is instantiated as a fixed-budget ordered list of normalized concept mentions mined directly from the segment transcript, rather than KnowledgeConcept IDs; E_i^{sub} is left empty when prerequisite edges are unavailable. ACP then operates on surface-form alias sets derived from the current mention string and treats π_i as the executable local order contract. Importantly, KPC and LOC remain offline evaluation-only signals and are never exposed as model inputs at test time.

4.5 Ablation Study

We conduct progressive ablation on the in-domain validation set (N=48, Table 4).

Planning Beats Static Retrieval. *Text-RAG (Top-K)* yields modest structural gains (KPC: 56.52→59.82, +5.8%), limited by absent trajectory planning. *SG-Planner* delivers the largest single-step jump (+2.28 to 62.10, $p < 0.01$), validating contextual pruning as the core enabler of prerequisite ordering.

Visual+Constraint Synergy. With the planner fixed, *KGVA-only* improves R1 and C-F1 and raises KPC from 62.10 to 63.22, indicating that plan-guided visual grounding mainly strengthens evidence alignment. *ACP-only* yields a larger KPC gain (62.10→64.43), consistent with its direct role in suppressing premature look-ahead under prerequisite-ordered decoding. Combining both gives the best Stage 2 result (KPC 65.60), suggesting complementary rather than redundant gains.

Global Coherence. *GKS* fusion attains peak performance (KPC 66.79), contributing 12% while reducing cross-segment discontinuities, consistent with its design goal.

Oracle Planning Diagnostic. To quantify error propagation from Stage 1, we replace the predicted plan with an oracle plan while keeping Stage 2 (KGVA+ACP) and Stage 3 (GKS) unchanged. On the in-domain Test ID split, oracle planning improves KPC from 66.79 to 70.20, LOC from 63.11 to 66.10, and ROUGE-L from 36.62 to 37.05, yielding gains of +3.41, +2.99, and +0.43, respectively. This gap suggests that residual structural errors are still dominated by Stage 1 planning quality, while the downstream generator can realize substantially better pedagogical structure when the plan is ideal. Detailed oracle construction and phase-level analyses are deferred to the Appendix.

4.6 Robustness Analysis

We evaluate robustness under ASR Noise (averaging ASR-Drop 30% and ASR-Corrupt 20%) and a strict Visual-Only setting where models receive only 8 frames without transcripts. For fairness, all methods use the same deterministic frame sampling and caching pipeline under the fixed visual budget. In the V-only setting, since SG-Planner requires textual input, we keep the same planning interface by generating dense visual captions via the frozen visual encoder to serve as surrogate queries, rather than bypassing Stage 1 altogether (details in Appendix A.6).

As shown in Table 5, the Full Model demonstrates superior stability, degrading by only 7.5% under transcript perturbations relative to in-domain performance. Even with visual-only input, it maintains 55.5 KPC, suggesting that plan-guided grounding and prerequisite-aware decoding provide a usable structural prior when textual evidence is unavailable. Similar trends were observed on

Method	Noise		V-only	
	KPC	LOC	KPC	C-F1
InternVL3	54.9	46.0	42.0	50.0
Text-RAG†	57.8	50.5	–	–
Full Model*	**61.9**	**59.3**	**55.5**	**60.5**

Table 5: **Robustness.** Noise: mean of ASR-Drop (30%) and ASR-Corrupt (20%). V-only: frames only under the same fixed visual budget; SG-Planner remains active by using VLM-generated surrogate queries. $^*p < 0.001$ vs. InternVL3 (paired Wilcoxon, video-level). † Text-RAG is not applicable to V-only.

the VideoLLaMA2 backbone, indicating that the robustness gain is not tied to a single backbone.

Efficiency Note: The lightweight planner adds negligible latency ($< 5\%$ overhead under the protocol in Appendix C.1), as end-to-end runtime is dominated by decoder generation. Additional stress tests under phase-boundary jitter and local interleaving perturbations are reported in the Appendix. For qualitative analysis of logical inversion correction and attention heatmaps, please refer to Appendix D.2 and Appendix D.5.

5 Conclusion

In this work, we formalize Structure-Pedagogical Reconstruction (SPR) to address the persistent issue of logical inversion in end-to-end Video-LLMs. We propose KCVR, a neuro-symbolic framework that strategically decouples evidence-grounded epistemic planning from topology-constrained realization. By synergizing dynamic graph pruning with the KGVA-ACP mechanism, KCVR enforces strict adherence to classroom prerequisite logic, effectively bridging the gap between perceptual retrieval and reasoning. Empirical evaluations on EduStruct substantiate that KCVR not only mitigates structural hallucinations but also demonstrates robust zero-shot transferability to graph-absent domains (e.g., InfoTech). Our findings suggest that for high-density knowledge scenarios, explicit structural priors provide a strong and reliable scaffold for trustworthy and interpretable generation. Future avenues will explore the joint optimization of planner-generator dynamics and extend this structure-centric paradigm to other high-stakes procedural domains, such as medical diagnostics.

Limitations

While KCVR establishes a new baseline for structured educational summarization, we acknowledge three limitations that outline directions for future research.

Dependence on Domain Priors. Our framework relies on a pre-defined Knowledge Graph (KG) to guide the SG-Planner. Although our **L3SO** experiments (Table 3) demonstrate that the model can generalize to graph-absent domains (e.g., InfoTech) by internalizing pedagogical meta-schemas, performance still drops by about 10.7% relative to the in-domain KPC under graph-supported settings (66.79→59.67 on InfoTech no-KG). This suggests that while KCVR is robust, it is not yet fully autonomous in "discovering" new knowledge structures from scratch. Future work could explore *Neural Graph Induction*, allowing the model to dynamically construct ephemeral KGs from the video stream itself during inference.

Pipeline Error Propagation. KCVR adopts a *Plan-then-Generate* pipeline rather than a fully end-to-end approach. While this explicit decomposition effectively prevents logical inversions, it introduces the risk of error propagation—for instance, if the SG-Planner prunes a critical prerequisite, the downstream generator cannot recover it. We mitigate this via the high-recall design of the planner ($B = 20$), but a joint optimization strategy (e.g., reinforcing the planner with generator feedback) remains a promising avenue to bridge the gap between discrete planning and continuous generation. An oracle planning diagnostic further shows that replacing the predicted plan with an oracle plan improves KPC by +3.41 and LOC by +2.99 under identical downstream settings, indicating that residual structural errors are still largely bottlenecked by Stage 1 planning quality.

Scope of Multimodal Interaction. Our current KGVA module focuses on grounding "static" visual evidence (e.g., blackboard formulas, slides) which is dominant in classroom settings. However, it may be less effective for highly dynamic, motion-centric instructional videos (e.g., physical education or laboratory experiments) where the pedagogical logic is embedded in temporal actions rather than symbolic text. Extending the SPR paradigm to capture procedural action logic is a vital next step for broader applicability.

Ethics Statement

We strictly adhere to the ACL Ethics Policy. This work involves the collection and computational analysis of real-world classroom videos, raising specific ethical considerations regarding privacy, bias, and potential misuse.

Privacy and Data Protection. The EduStruct dataset comprises recordings collected from teaching competitions with explicit informed consent from all participating instructors and institutional approval from the hosting university's ethics committee. Although the videos were originally public within the competition scope, we implement a rigorous **De-identification Protocol** for research release (detailed in Appendix A.8). This includes (1) automated detection and blurring of faces for both teachers and students, and (2) scrubbing of Personally Identifiable Information (PII) such as school names and spoken names from transcripts and metadata. The released artifacts are strictly limited to derived features and annotations, preventing the reconstruction of biometric identities.

Curriculum and Algorithmic Bias. We acknowledge that the Knowledge Graphs (KGs) driving our system are derived from standardized national curriculum textbooks. While we perform canonicalization to abstract away publisher-specific biases, the graphs may still reflect the pedagogical sequencing preferences of the source region. However, since KCVR focuses on extracting *structural logic* (e.g., mathematical derivations) rather than subjective narratives, the core epistemic dependencies remain largely universally applicable. Users deploying KCVR in humanities subjects should be aware of potential source-level narrative biases.

Intended Use vs. Misuse. The intended application of KCVR is to facilitate post-hoc knowledge retrieval for students and lesson planning assistance for educators. We explicitly condemn the use of this technology for automated surveillance or punitive evaluation of teacher performance (e.g., using "Logical Inversion" rates to penalize teachers). The metrics proposed in this work are designed to evaluate *summarization models*, not *human instructors*. We urge practitioners to prioritize teacher autonomy and use such tools solely for supportive scaffolding.

Acknowledgment

We acknowledge the support of the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2026C02A1236) and the National Natural Science Foundation of China under Grant (No. 62577050).

References

Kevin Ackermans, Björn B. de Koning, and Halszka Jarodzka. 2025. Instructional videos and deeper processing: Insights and applications. *Learning and Instruction*, 98:102137.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3278–3292.

Fatemeh Bagheri, Anshuman Garga, and Ramon E. Lopez. 2024. Exploring radio emissions from confirmed exoplanets using SKA. *Preprint*, arXiv:2404.14468.

Z. Chen and 1 others. 2024. Intern VL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Lishuai Gao, Yujie Zhong, Yingsen Zeng, Haoxian Tan, Dengjie Li, and Zheng Zhao. 2024. Linvt: Empower your image-level large language model to understand videos. *Preprint*, arXiv:2412.05185.

Hanghui Guo, Jia Zhu, Shimin Di, Weijie Shi, Zhangze Chen, and Jiajie Xu. 2025. DioR: Adaptive cognitive detection and contextual retrieval optimization for dynamic retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2953–2975.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025. Retrieval-augmented generation with graphs (GraphRAG). *Preprint*, arXiv:2501.00309.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and

Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *Preprint*, arXiv:2404.05726.

Weifeng Hu, Yu Zhang, Yujun Li, Jia Zhao, Xifeng Hu, Yan Cui, and Xuejing Wang. 2023. Query-based video summarization with multi-label classification network. *Multimedia Tools Appl.*, 82(24):37529–37549.

Hang Hua, Yolo Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2025. V2Xum-LLM: Cross-modal video summarization with temporal prompt instruction tuning. *Preprint*, arXiv:2404.12353.

Zhisheng Huang, Xudong Jia, Tao Chen, and Zhongwei Zhang. 2025. A general scalable approach for knowledge selection based on iterative hybrid encoding and re-ranking. *Data Intelligence*, 7(1):124–142.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. *Preprint*, arXiv:2311.08046.

Prem Nigam Kar, David E. Roberson, Tim Seppelt, and Peter Zeman. 2025. NPA hierarchy for quantum isomorphism and homomorphism indistinguishability. *Leibniz International Proceedings in Informatics*, 334:105:1–105:19.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. *Preprint*, arXiv:2106.10502.

Shuailin Li, Yuang Zhang, Yucheng Zhao, Qiuyue Wang, Fan Jia, Yingfei Liu, and Tiancai Wang. 2023. Vlm-eval: A general evaluation on video large language models. *Preprint*, arXiv:2311.11865.

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 Transactions on Multimedia*, 26:5548–5560.

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.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2022. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States. Association for Computational Linguistics.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. 2025. Graph-constrained Reasoning: Faithful reasoning on knowledge graphs with large language models. *Preprint*, arXiv:2410.13080.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. 2023. Pg-video-llava: Pixel grounding large video-language models. *Preprint*, arXiv:2311.13435.

Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. CLIP-It!: Language-guided video summarization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Torbjørn Netland, Oliver von Dzengelevski, Katalin Tesch, and Daniel Kwasnitschka. 2025. Comparing human-made and AI-generated teaching videos: An experimental study on learning effects. *Comput. Educ.*, 224(C).

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From dense token to sparse memory for long video understanding. *Preprint*, arXiv:2307.16449.

Yolo Y. Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang

Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2025. Video understanding with large language models: A survey. *Preprint*, arXiv:2312.17432.

Lifu Tu, Semih Yavuz, Jin Qu, Jiacheng Xu, Rui Meng, Caiming Xiong, and Yingbo Zhou. 2024. Unlocking anticipatory text generation: A constrained approach for large language models decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15532–15548, Miami, Florida, USA. Association for Computational Linguistics.

Shuo Wang and Jihao Zhang. 2025. MF2Summ: Multimodal fusion for video summarization with temporal alignment. *Preprint*, arXiv:2506.10430.

Peng Xiao, Chao Liu, Wei Jia, and Lijun Dong. 2025. Aligned-entities-based fusion embedding on hetero-field knowledge graphs. *Data Intelligence*, 7(3):618–635.

Tao Xu, Yuan Liu, Yaru Jin, Yueyao Qu, Jie Bai, Wenlan Zhang, and Yun Zhou. 2025. From recorded to AI-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology*, 56(4):1463–1487.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. *Preprint*, arXiv:2406.16620.

Jia Zhu, Hanghui Guo, Weijie Shi, Zhangze Chen, and Pasquale De Meo. 2025. RaDIO: Real-time hallucination detection with contextual index optimized query formulation for dynamic retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26129–26137.

Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. 2021. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962.

A Data and Annotation Guidelines

Overview and notation. EduStruct is organized at two aligned granularities, with a strict contract that separates model inputs, supervision signals, and evaluation-only metadata. At the *video level*, each lesson provides one macro annotation file `*_global.json`, which contains the global structured summary (`summary_edu_gt`) and video-level learning objectives (`learning_objectives`). At the *phase level*, each lesson is segmented into four pedagogical phases (*Introduction*, *Exposition*, *Interaction*, *Conclusion*) and each phase has a dedicated JSON file containing the phase transcript (`transcript`), the phase target summary (`summary_gt`), phase-level learning objectives, and `knowledge_points` that are aligned to KnowledgeConcept nodes in the dual-layer epistemic graph.

Phase boundaries are obtained by a transcript-only pedagogical phase parser; to prevent leakage, phase parsing is trained and selected using only in-domain training and validation splits, with constraints enforced at the video-id level. Our experiments follow this hierarchy end-to-end: the generator is trained and evaluated on phase-level micro-generation, and macro reconstruction is produced by fusing four phase summaries with the Global Knowledge Sketch (GKS), as detailed in Appendix A.5. To make supervision boundaries explicit, Appendix A.3 enumerates, for every field in the released JSONs, whether it is used as model input, as supervision, or as evaluation-only metadata (e.g., learning objectives and knowledge points for LOC and KPC computation). All split and leakage constraints are enforced at the video-id level, and an audit script verifies that no video appears in more than one split; additional annotation consistency checks and the mapping to the dual-layer epistemic graph are documented in Appendix A.4.

A.1 Inclusion and exclusion criteria

We apply deterministic inclusion and exclusion criteria to ensure that each lesson supports phase-level supervision and that all methods are evaluated under identical modality budgets and preprocessing.

Inclusion criteria. A lesson video is included if it satisfies all conditions below: (1) The lesson admits four pedagogical phases (*Introduction*, *Exposition*, *Interaction*, *Conclusion*) with valid timestamps, and each phase has a non-empty transcript. (2) Each phase annotation contains a phase-level target summary and at least one learning objective. Learning objectives are part of the held-out evaluation contract for computing LOC and are not used as test-time inputs; in the objective-aware variant, they are used only as training supervision. (3) For each phase segment, frames can be deterministically sampled and cached under the fixed visual budget used throughout the paper (8 frames per segment). The same sampling policy and caching pipeline are shared across all methods (including baselines) to ensure strict modality parity.

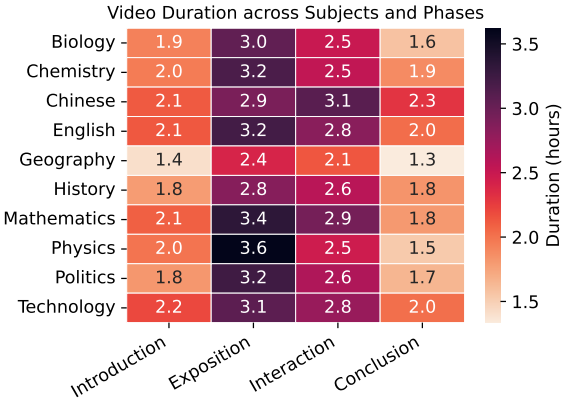

Figure 4: **Phase-level duration profile across subjects.** Heatmap of total video duration (hours) aggregated by subject (rows) and pedagogical phase (columns: Introduction, Exposition, Interaction, Conclusion). This phase-level distribution characterizes the evidence density of micro generation and complements the overall subject distribution shown in Fig. 3.

Exclusion criteria. We exclude lesson videos that violate any of the following: (1) Missing audio or severe ASR failure such that one or more phase transcripts are empty or largely unintelligible. (2) Annotation incompleteness, including missing phase targets or missing learning objectives in any phase. (3) Severe visual corruption (e.g., overexposure, extreme blur, or full occlusion) that prevents consistent frame sampling or reliable visually grounded evidence from blackboard or slide content. (4) Privacy-sensitive content that cannot be reliably de-identified under our dataset release policy.

Reproducibility note. To ensure that all dataset statistics reported in Appendix A use a single source of truth, we export three derived files from the released annotations: (i) `video_stats.csv` (video id, subject, duration, split), (ii) `phase_stats.csv` (video id, phase label, timestamps, transcript length, objective and

concept counts), and (iii) `subject_stats.csv` (aggregated subject-level counts and durations). All numbers in Fig. 3 and Fig. 4 are computed from these files using the same script.

A.2 Phase Segmentation Protocol and P^3 Implementation

EduStruct uses a fixed four-phase pedagogical protocol (*Introduction*, *Exposition*, *Interaction*, *Conclusion*) as the atomic unit for micro-supervision. This section specifies the segmentation contract and the transcript-only Pedagogical Phase Parser (P^3).

Four-phase contract. The four-phase schema follows the dominant lesson-plan structure in our source data and provides a consistent unit for phase-level targets, learning objectives, and concept alignment. We standardize all lessons into this contract so that each phase yields a segment (T_i, I_i) for Stage 1 planning and Stage 2 micro-generation.

P^3 and inference interface. P^3 is a transcript-only sequence labeling model that assigns one of four phase labels to each utterance. At inference time, P^3 takes only the raw transcript and produces a phase label sequence; it does not access learning objectives, gold phase targets, or any test-time annotations.

From labels to temporal boundaries. We deterministically convert utterance labels into temporal segments as follows. Consecutive utterances with the same predicted phase label are merged into a maximal contiguous block; each block defines one phase span. The start time of a phase span is the timestamp of the first utterance in the block, and the end time is the timestamp of the last utterance in the block; the corresponding transcript slice defines T_i and frames are sampled within the span to form I_i under the fixed visual budget used in the main experiments. If a phase label is missing due to prediction errors, we keep the resulting spans as-is rather than forcing a hard four-block post-hoc correction, and rely on downstream robustness analyses reported in this appendix.

Boundary evaluation. We report boundary quality on phase transition points derived from utterance labels. Let $\mathcal{B}_{\text{gold}}$ and $\mathcal{B}_{\text{pred}}$ be the sets of gold and predicted boundary timestamps. A predicted boundary is counted as correct if it can be matched to an unmatched gold boundary within a fixed tolerance window τ seconds using nearest-neighbor matching; unmatched predicted boundaries are false positives and unmatched gold boundaries are false negatives. We compute boundary F1 from the matched boundaries, and report the median absolute offset (in seconds) over matched pairs.

Training split and leakage control. P^3 is trained only on Train (ID) and selected using Val (ID). No video from Test (ID) or any OOD test subject is used to train or select P^3, enforced at the video-id level. Performance details (Phase Boundary F1=82.3%, median offset=14s), the tolerance τ, and implementation settings are provided in this appendix.

A.3 Data splits, OOD protocol, and leakage control

Protocol Definitions We evaluate structural generalization via a *Leave-Three-Subjects-Out (L3SO)* protocol (Table 7).

- **In-Domain (ID)**: 7 STEM and Humanities subjects used for training and validation.
- **KG-Supported OOD** (History, English): Held-out subjects with KGs available at inference. Tests **Epistemic Transfer** across distinct knowledge types.
- **KG-withheld OOD** (InfoTech): A strictly adversarial setting with **NO** Knowledge Graph at inference. Crucially, InfoTech represents a **Double OOD** challenge:
  1. **Structural OOD**: The model must infer structure without graph guidance.
  2. **Visual-Semantic OOD**: While recorded in the same classroom setting, the visual focus shifts from *Handwritten Blackboard Derivations* (typical in ID subjects) to ***Dense Code Projections*** and IDE interfaces. The epistemic logic also shifts from *Declarative Knowledge* to ***Procedural Logic*** (e.g., debugging loops).

  This setting rigorously tests whether KCVR has internalized universal pedagogical meta-schemas that persist even when the domain content and visual modality undergo significant shifts.

Split	Videos	Phase segments	Avg seg/video
Train (ID)	224	896	4
Val (ID)	48	192	4
Test (ID)	48	192	4
Test (OOD)	143	572	4
Total	463	1852	4

Table 6: Video-level splits for summarization under the fixed four-phase protocol.

Subject	Role in L3SO	KG@Eval
Math	ID (Train/Val/Test)	✓
Chinese	ID (Train/Val/Test)	✓
Physics	ID (Train/Val/Test)	✓
Chemistry	ID (Train/Val/Test)	✓
Politics	ID (Train/Val/Test)	✓
Biology	ID (Train/Val/Test)	✓
Geography	ID (Train/Val/Test)	✓
History	OOD Test	✓
English	OOD Test	✓
InfoTech	OOD Test (KG-withheld)	×

Table 7: Subject roles in L3SO and KG availability at evaluation time.

Key distinction (subject vs. video splits). L3SO is defined at the *subject* level: OOD subjects (History/English/InfoTech) contribute videos only to **Test (OOD)**. Train (ID), Val (ID), and Test (ID) are *video-level* splits drawn from the remaining 7 ID subjects.

Summarization splits All split constraints are enforced at the video-id level, and all four phase segments from the same video stay in the same split. The overall split statistics under the fixed four-phase protocol are:

Subject roles and KG availability To make the L3SO protocol auditable, Table 7 summarizes each subject's role (used for training, validation, testing) and whether its subject knowledge graph is available at evaluation time. In particular, the three OOD subjects are excluded from summarization training and validation, and InfoTech is evaluated under a no-KG condition.

P^3 training and leakage control To prevent leakage through segmentation, P^3 is trained only on Train (ID). All hyperparameter choices and early stopping for P^3 are decided using Val (ID) only. No video from the summarization test split is used to train or select P^3, including both Test (ID) and all OOD test subjects, and this constraint is enforced at the video-id level.

Evaluation-only signals. Learning objectives and knowledge points are used *only* for offline computation of KPC/LOC and are never exposed as model inputs; this evaluation contract remains available even when the subject KG is withheld at inference (e.g., InfoTech), and thus does not imply any test-time knowledge injection.

KG-withheld inference interface. Under the strict KG-withheld setting (InfoTech), Stage 1 keeps the same planning interface $P_i = (V_i^{sub}, E_i^{sub}, \pi_i)$ required by Stage 2 and Stage 3, but degenerates from KG-based subgraph planning into a plan-induction mode without accessing any external subject graph at inference time. Concretely, V_i^{sub} is instantiated as a fixed-budget ordered list of normalized concept mentions mined directly from the segment transcript, rather than KnowledgeConcept IDs. We apply deterministic normalization including Unicode canonicalization, lowercasing, whitespace collapsing, punctuation normalization, and simple CamelCase splitting, and keep the same planner budget $B = 20$. Since no external prerequisite edges are available, E_i^{sub} can be left empty in this setting, while π_i serves as the executable local order contract for downstream decoding. ACP then operates on surface-form alias sets derived from the current mention string and suppresses aliases of strictly future mentions under π_i. This protocol preserves the inference interface while avoiding any test-time injection of external prerequisite structure.

A.4 Annotation consistency and quality control

Annotator Calibration and Audit EduStruct annotations are produced by a team with education-related graduate training, assisted by subject-specific teaching consultants. All dataset annotators were compensated at rates meeting or exceeding the local minimum wage. Before full-scale annotation, annotators complete a calibration stage on a shared subset to align granularity and terminology, and disagreements are used to update the written guidelines. During production, each sample is annotated by two independent annotators. In addition, we perform a 10% spot-check audit by a second reviewer, and unresolved cases are escalated to a senior reviewer.

Each video is segmented into four pedagogical phases, and annotations are provided at both the segment level and the full-video level. For each seg-

ment, the dataset includes learning objectives and knowledge points as structured fields, in addition to transcript and summaries, enabling phase-aware evaluation of pedagogical quality. These fields are used to compute LOC and KPC, and therefore are verified with an explicit mapping rule to the curriculum graph.

Mapping to TeachingPoint and KnowledgeConcept. EduStruct uses a dual-layer epistemic graph where TeachingPoint represents curriculum units and KnowledgeConcept represents atomic terms and concepts. In our annotation schema, learning objectives are aligned to the TeachingPoint layer when they describe a curriculum unit, while knowledge points are aligned primarily to the KnowledgeConcept layer to support dependency-based evaluation. When a text span matches a KnowledgeConcept name (or its normalized alias), it is mapped directly to that node; otherwise, it is first mapped to a TeachingPoint and then expanded to candidate concepts through HAS_CONCEPT links, followed by reviewer confirmation. KPC is computed over prerequisite relations defined on the KnowledgeConcept layer (e.g., BASED_ON), while LOC evaluates whether learning objectives are covered by the generated summary under the same annotation contract. Aliases are normalized by Unicode canonicalization, case folding, and domain-specific symbol normalization (e.g., formula variants), with a subject-specific alias lexicon released with the KG. For TP-first cases, we expand to the Top-n KCs under HAS_CONCEPT (default $n = 10$) and require the reviewer to select the best-matching KC or mark it as out-of-scope.

A.5 Phase-Level Micro-Generation and Evaluation

Micro-segment scale Our generator is trained and evaluated at the micro level on four pedagogical phases (Introduction, Exposition, Interaction, Conclusion), and the macro summary is obtained by fusing phase-wise outputs with the Global Knowledge Sketch (GKS). Since each video is deterministically partitioned into four phases, the number of micro segments equals four times the number of videos in each split.

Objectives and leakage boundary. Learning objectives and knowledge points are part of the held-out annotation contract for evaluation (LOC/KPC) and are never provided as test-time inputs for the main model. For the optional objective-aware variant (Appendix E.3), we use objectives only as supervision during training; at inference, any objective signal used for conditioning is self-generated by the model from the same (T,V) evidence (Chain-of-Thought), and no gold objectives from the dataset are exposed.

Phase	Train (ID)	Val (ID)	Test (ID)
Introduction	224	48	48
Exposition	224	48	48
Interaction	224	48	48
Conclusion	224	48	48
Total	896	192	192

Table 8: Micro-segment counts per phase for in-domain training and evaluation. Counts are computed as videos $\times$ 4 under the fixed four-phase protocol.

Method	Phase	R-L	C-F1	KPC	LOC
SFT Baseline	Introduction	33.8	64.0	58.2	56.1
	Exposition	34.6	65.1	57.4	55.3
	Interaction	32.9	63.2	54.6	52.8
	Conclusion	34.1	65.0	56.7	55.0
Text-RAG (Top-K)	Introduction	34.2	66.5	61.1	59.0
	Exposition	34.9	67.6	60.3	58.2
	Interaction	33.4	66.8	57.2	55.7
	Conclusion	34.5	67.7	60.0	58.4
Full Model (Ours)	Introduction	36.8	70.1	69.0	65.0
	Exposition	37.1	71.2	67.5	64.1
	Interaction	35.4	69.4	64.2	60.8
	Conclusion	36.9	70.8	66.5	62.7

Table 9: Phase-wise micro-summary results on the in-domain test split. Interaction is consistently the most challenging phase due to open-ended Q&A and error correction, where implicit prerequisites are frequent.

Phase-wise micro results Table 9 reports phase-wise micro-summary performance on the in-domain test split. Compared with macro-only reporting, this breakdown validates that the micro generator learns phase-specific behaviors, and it also localizes where the gains of planning and visual grounding originate.

A.6 Noise modeling rationale for robustness evaluation

Motivation and realism of perturbations Classroom transcripts obtained from far-field microphones are frequently incomplete or corrupted due to overlapping student speech, teacher movement, and domain-specific terminology. To approximate these failure modes in a controlled manner, we adopt two transcript perturbations. ASR-Drop randomly removes 30% utterances to simulate miss-

ing transcript spans, while ASR-Corrupt replaces 20% words using the top-1000 substitution patterns mined from real ASR logs, reflecting systematic confusions rather than arbitrary noise. All methods are evaluated under identical visual budgets (8 frames per segment) to ensure the robustness gains are not caused by extra visual context. For fairness, we pre-generate a perturbed transcript for each segment with a fixed random seed and reuse it across all methods. The substitution dictionary for ASR-Corrupt is mined from a disjoint ASR log pool and does not use any transcripts from Val/Test splits.

Visual-Only Implementation Details We evaluate transcript-free (V-only) inference to test reliance on blackboard evidence. Since the SG-Planner inherently requires textual queries, in the V-only setting we keep the same planning interface by generating dense visual captions as surrogate queries, rather than bypassing Stage 1 altogether. **Implementation**: We use the same frozen visual encoder as the backbone model (e.g., InternVL3) with a prompt such as *"Describe the blackboard content and teacher's key gestures in detail."* The generated caption (e.g., *"The teacher writes $F = ma$ on the board..."*) serves as the surrogate query for the Planner. All methods use the same deterministic frame sampling and caching pipeline under the fixed visual budget of 8 frames per segment, so the V-only comparison does not introduce extra visual evidence. This keeps the pipeline operational even without audio and isolates the contribution of plan-guided visual grounding under transcript-free input. Failures in this setting typically correspond to sparse blackboard content or occlusion, directly linking performance to visual evidence quality. We provide the exact prompt template for this query generation in Appendix F.

Stress tests under phase perturbation. To probe robustness beyond transcript corruption, we further test the pipeline under two controlled perturbations of the phase segmentation while keeping the backbone, decoding settings, and total visual budget fixed. For boundary jitter, we shift predicted phase boundaries by ± 15s before re-running Stage 1 to Stage 3 end-to-end. For local interleaving, we swap 20 utterances in a small window around the Exposition-Interaction boundary and reconstruct phase spans using the same deterministic span-construction rule. Since these perturbations break alignment with phase-level targets, we evaluate only the final macro summary.

Setting	KPC	LOC	ROUGE-L
Original spans	66.79	63.11	36.62
Boundary jitter ($\pm$15s)	65.52	62.06	36.31
Local interleaving	64.87	61.55	36.29

Table 10: Robustness under controlled phase perturbations on Test (ID) macro summaries.

A.7 Cross-Disciplinary Structural Diversity

To demonstrate that **EduStruct** captures the authentic epistemic logic across diverse disciplines, rather than merely fitting domain-specific heuristics (e.g., mathematical derivations), we present a qualitative montage in Figure 5. Each panel visualizes a micro-segment sample as a structured tuple $(V, T, \mathcal{O}, \mathcal{K})$, comprising:

- **Visual Evidence** (V): A representative keyframe (e.g., blackboard formulas, slides) showing the instructional focus.
- **Transcript Snippet** (T): The raw audio transcription, often colloquial and unstructured.
- **Learning Objective** ($\mathcal{O}$): The pedagogical intent (e.g., *Master the syntax*), serving as the high-level goal.
- **Knowledge Point** ($\mathcal{K}$): A knowledge point (KP) aligned to *KnowledgeConcept* in the dual-layer epistemic graph, grounding the segment to a concept-level node (e.g., *Loop Control*).

Epistemic Breadth. The montage highlights three distinct structural typologies covered in our benchmark: (1) **Symbolic Reasoning** (Math, Physics, Chemistry): Heavily reliant on blackboard derivations and rigorous definitions. (2) **Causal & Contextual Logic** (History, Politics, English): Focused on narrative arcs, historical causality, and language contexts. (3) **Procedural Logic** (InfoTech): Demonstrated by the *Python Loops* example, which requires understanding code execution flow, a key factor in our OOD adversarial testing.

A.8 Dataset Release & Ethical Considerations

Provenance and consent. EduStruct comprises 463 classroom videos recorded during internal teaching competitions at a normal university. The instructors are pre-service or in-service teachers

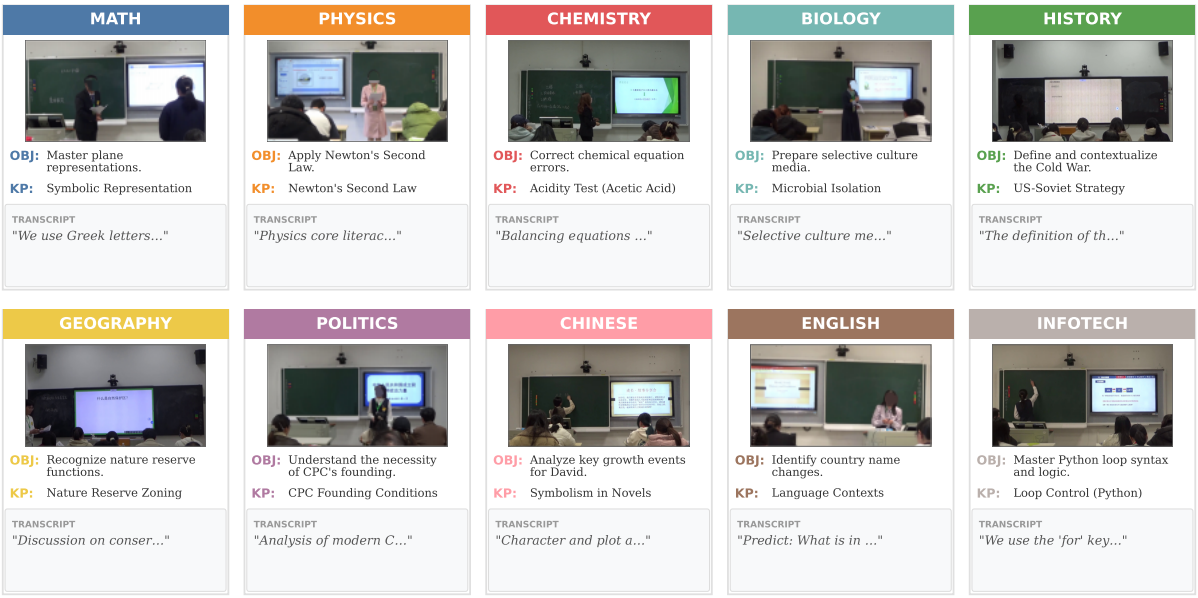

Figure 5: **Cross-subject qualitative montage.** We visualize 10 representative micro-segments from EduStruct. Note the rigorous distinction between the abstract **Knowledge Point (KP, KnowledgeConcept-aligned)** and the concrete **Learning Objective (OBJ)**, as well as the diversity of visual evidence (from chemical equations to Python code). This structural granularity enables KCVR to distinguish pedagogical intent from surface content.

with at least one year of teaching experience, and the recordings were originally created for instructional evaluation and teacher training under institutional permission. All participants provided explicit informed consent for research usage, and the collection and processing pipeline was reviewed and approved by the host institution's ethics committee.

Release scope and boundary. We do not distribute raw videos in the public release. The released package is limited to derived annotations and resources for research reproducibility (phase boundaries, transcripts, learning objectives, knowledge points, subject graphs, and evaluation scripts), with de-identification applied to all textual artifacts. When the corresponding teaching material is already publicly accessible, we additionally provide optional public clip indices to facilitate verification. These indices only point to pre-existing public sources (when available) and do not redistribute any video content from EduStruct. Raw videos may be accessed only via a controlled institutional request process, subject to privacy review.

De-identification protocol. All released artifacts are de-identified to mitigate privacy risks. For visual content, we automatically detect and blur/crop faces, school names or logos, and identifiable signage. For textual content, we scrub Personal Identifiable Information (PII) from transcripts and annotations, including names of teachers, students, or schools. Only pedagogical information necessary for structure analysis (e.g., blackboard content and slide text) is retained, and no identity-bearing metadata is included in the release package.

B Dual Layer Epistemic Graph and Schema

Across 10 subjects, the dual-layer epistemic graphs exhibit a stable scale and composition (Fig. 6). Each subject contains roughly 100–250 TeachingPoint (TP) nodes and 600–1000 KnowledgeConcept (KC) nodes, with 2.0k–3.8k typed edges per subject (1.5k TP, 8.5k KC, 30.5k relations in total). Relation distributions are consistent with the intended schema: PART_OF/NEXT form the curriculum backbone on the TP layer, HAS_CONCEPT and RELATED_TO dominate cross-layer and lateral connectivity, while BASED_ON and the sparse PREREQ_OF edges provide directed prerequisite signals exploited by SG-Planner.

B.1 Dual-Layer Epistemic Graph Design

We formally define a subject-specific epistemic graph as $G^{(s)} = (V_{TP}^{(s)} \cup V_{KC}^{(s)}, E_{TP}^{(s)} \cup E_{KC}^{(s)} \cup E_{A}^{(s)})$. Our core design philosophy is to decouple *narrative scope* from *epistemic logic*: the TeachingPoint (TP) layer constrains **what** is discussed, while the

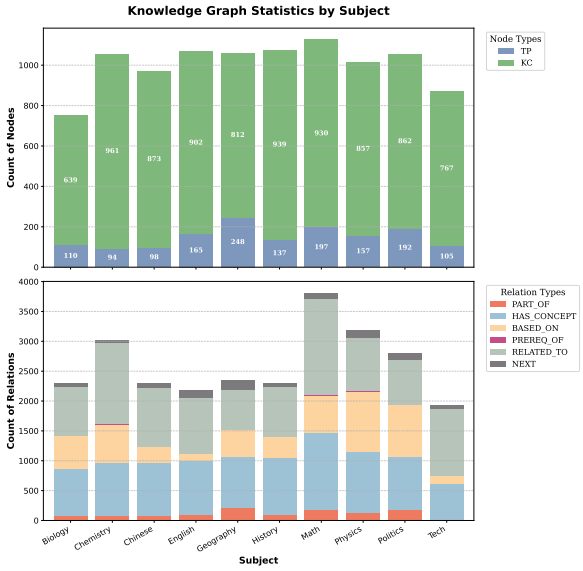

Figure 6: **Dual-layer epistemic graph statistics across subjects.** Top: TP and KC node counts per subject. Bottom: edge counts by relation type (PART_OF, HAS_CONCEPT, BASED_ON, PREREQ_OF, RELATED_TO, NEXT), showing a TP-backbone plus KC connectivity pattern used by SG-Planner.

KnowledgeConcept (KC) layer governs **how** concepts are ordered.

Standardization and Canonicalization. To ensure that our graph captures universal pedagogical structures rather than overfitting to a specific textbook version (e.g., regional editions)(Xiao et al., 2025), we applied a **Curriculum-Agnostic Canonicalization** process. Specifically, the **TeachingPoint** layer is constructed by abstracting the intersection of curriculum standards from multiple authoritative textbooks. We normalize node identifiers (e.g., utilizing semantic codes like `tp_math_func_deriv` instead of book-specific indices) to remove publisher-specific artifacts. Crucially, the **KnowledgeConcept** layer encodes scientific axioms and logical dependencies (e.g., *Limit* $\rightarrow$ *Derivative*) that are **invariant across languages and educational systems**. This ensures that the inferred prerequisites reflect universal epistemic truth rather than arbitrary curricular sequencing, enabling EduStruct to serve as a generalizable benchmark for logic-driven instructional summarization.

Curriculum Layer: TeachingPoint (TP). TP nodes encapsulate the textbook narrative hierarchy. Relations PART_OF and NEXT jointly define a stable curriculum backbone. This structure prevents the planner from drifting into out-of-scope content by strictly bounding the instructional context within the parent unit hierarchy.

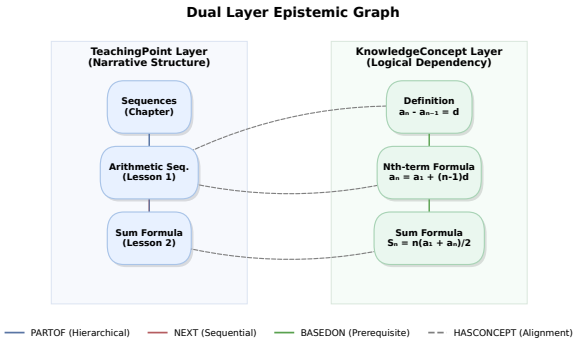

Figure 7: **Math Subgraph Topology.** Left: TP layer showing curriculum hierarchy. Right: KC layer showing epistemic dependencies. Cross-layer alignment enables scope-constrained planning.

Epistemic Layer: KnowledgeConcept (KC). KC nodes represent atomic, reusable concepts. BASED_ON forms the dense dependency backbone, capturing fine-grained derivational logic (e.g., Definition $\rightarrow$ Formula). Complementing this, PREREQ_OF acts as a sparse, high-penalty constraint marker for critical pedagogical jumps. In our SG-Planner, BASED_ON provides general precedence signals, while PREREQ_OF enforces hard topological constraints to prevent severe logical inversions.

Cross-Layer Alignment. HAS_CONCEPT edges bridge the narrative and epistemic layers (E_A). This alignment enables a two-step planning mechanism: first filtering candidates by TP scope, then linearizing them via KC dependencies.

Case Study: Arithmetic Sequences (Math). Figure 7 illustrates this topology. The TP layer (left) dictates the lesson sequence ("Lesson 1: Definition" $\rightarrow$ "Lesson 2: Summation"), while the KC layer (right) enforces the logical derivation ("Definition" $\xrightarrow{\text{BASED\_ON}}$ "Nth-term Formula" $\xrightarrow{\text{BASED\_ON}}$ "Sum Formula"). The SG-Planner leverages this dual structure to retrieve scope-constrained prerequisites.

B.2 Universal Node and Relation Schema

To facilitate cross-disciplinary and **cross-curriculum** alignment, all subject graphs adhere to a unified schema definition (Table 11). This standardization ensures that the SG-Planner allows the KCVR framework to operate agnostically across domains **and educational standards (e.g., K-12 vs. International Baccalaureate).**

Element	Semantics and Structural Constraints
Node Types	
TeachingPoint (TP)	Represents *Narrative Units* (e.g., Chapter, Lesson). *Constraints:* Strict hierarchy ($\texttt{level} \in \{1, 2, 3\}$); unique ID per subject. Serves as the curriculum anchor.
KnowledgeConcept (KC)	Represents *Epistemic Atoms* (e.g., Theorems, Terms). *Constraints:* Must contain a non-empty textual `definition` for embedding injection.
Edge Types	
PART_OF (TP→TP)	Hierarchical containment (Chapter ⊃ Section). *Constraint:* $\texttt{level}_{start} < \texttt{level}_{end}$. Enforces tree structure.
NEXT (TP→TP)	Pedagogical sequence within a parent unit. *Constraint:* Typically intra-sibling links only.
HAS_CONCEPT (TP→KC)	Cross-layer alignment bridge. *Constraint:* Many-to-many allowed; enables scope-constrained retrieval.
BASED_ON (KC→KC)	Dense dependency backbone (Derivational/Definitional). *Constraint:* Strictly acyclic (DAG) for STEM subjects.
PREREQ_OF (KC→KC)	Sparse, high-penalty prerequisite marker. *Constraint:* No self-loops; defines hard ordering barriers.

Table 11: Universal schema definition shared across all 10 subjects. The separation of TP (Narrative) and KC (Epistemic) layers is the architectural foundation of our KCVR framework.

B.3 Graph Quality Assurance Checklist

We enforce rigorous quality control to ensure topological validity. Table 12 details the automated audit protocols applied to each subject graph before deployment.

C SG-Planner Details and Planner-Level Evidence

This appendix provides controlled evidence for SG-Planner's internal validity, addressing three questions: (i) sensitivity to pruning budget B, (ii) the isolated effect of minimum-violation ordering, and (iii) correlation between planner metrics and downstream KPC/LOC. The planner outputs $\mathcal{P} = (V_{sub}, E_{sub}, \pi)$ for each segment, where V_{sub} is the budgeted concept set, E_{sub} the induced edges, and π the topological traversal path.

Audit Metric	Validation Protocol & Pass Criteria
ID Uniqueness	Global uniqueness verification across node sets. *Pass:* $\lvert V\rvert = \lvert unique(IDs)\rvert$.
DAG Topology	Cycle detection on BASED_ON edges. *Pass:* No cycles in STEM; minimal feedback loops ($< 1\%$) in Humanities (resolved by Min-Violation Ordering during planning).
Hierarchy Check	Tree validity check for PART_OF relations. *Pass:* Max in-degree ≤ 1; strictly increasing level depth.
Concept Alignment	Coverage analysis of HAS_CONCEPT edges. *Pass:* $> 80\%$ of leaf TPs have ≥ 1 aligned KC.
Metadata Integrity	Null-value audit for textual descriptions. *Pass:* 0% missing definitions; avg. length > 20 chars.
Subject Isolation	Partition integrity check. *Pass:* No cross-subject edges; graphs are strictly disjoint components.

Table 12: Automated graph quality checklist. All 10 subject graphs passed these validations.

C.1 SG-Planner Configuration

SG-Planner uses expansion hop $h = 2$ and pruning budget $B = 20$ (main setting), fixed before evaluation. Evidence scores s_{tp}, s_{kc} are computed via dense retrieval (embeddings of name/definition) against the phase transcript. Note that OCR tokens are used **only in the specific OCR-analysis setting** (Appendix E.4); the main setting relies solely on transcript and KGVA for grounding.

The curriculum anchor v^*_{tp} defines the local scope Φ via PART_OF/NEXT edges. The candidate set U_{cand} is built by h-hop expansion from top-K concepts by s_{kc} (primarily along BASED_ON). Pruning selects V_{sub} by maximizing $s_{kc}(u) + \lambda \cdot \mathcal{C}(u)$, where $\mathcal{C}(u)$ is the curriculum prior derived from HAS_CONCEPT edges (soft bias λ favors in-scope concepts).

Efficiency note. With $B \leq 20$ and a greedy adjacent-swap local search, the ordering step runs in $O(B^2 \cdot |E^{\rightarrow}_{kc}|)$ per segment in practice. Under this fixed configuration, the planner contributes $< 5\%$ **end-to-end latency overhead** under the measurement protocol referenced in Sec. 4.6. End-to-end runtime is dominated by decoder generation, which explains the small planner ratio despite explicit retrieval, pruning, and ordering.

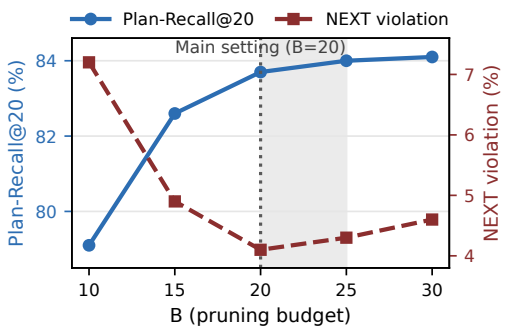

Figure 8: Budget sensitivity of SG-Planner. Plan-Recall@20 (left axis) and NEXT violation rate (right axis) vs. B. Shaded region: stability plateau. Main setting $B = 20$ (dashed line).

C.2 Budget Sensitivity Analysis

Figure 8 shows the trade-off between coverage (Plan-Recall@20) and coherence (NEXT violation) as B varies. A stable plateau emerges for $B \in [20, 25]$, where Plan-Recall@20 nears saturation ($\approx 83.7\%$) while NEXT violations remain minimal ($\approx 4.1\%$). The BASED_ON violation rate is consistently low ($< 4.5\%$) across this range. We select $B = 20$ as it lies at this plateau's elbow without over-tuning for any single metric.

C.3 Isolating the Effect of Minimum-Violation Ordering

To disentangle ordering from node selection, we fix V_{sub} ($B = 20$) and compare: (i) **Init**: sorting by descending relevance score s_{kc}, and (ii) **Refined**: applying greedy local swaps (Alg. 1) to minimize violations. Specifically, we initialize π by descending s_{kc}, then iteratively perform adjacent swaps that reduce the violation cost $\sum w_{u,u'}$, terminating when no local improvement is possible.

Figure 9 visually isolates this effect. The initial score-based ranking results in a tangled dependency graph with frequent logical inversions (red arcs in Fig. 9a), as retrieval scores often bias towards high-frequency terms like "Summation" over prerequisites like "Definition". In contrast, our refined ordering untangles these dependencies, linearizing the pedagogical flow (blue arcs in Fig. 9b). Quantitatively, this process reduces total violations from 11.2% to 8.4% (identical V_{sub}), with the strongest gain in NEXT edges (local teaching sequence: -42%), confirming that ordering recovers pedagogical linearity beyond relevance ranking.

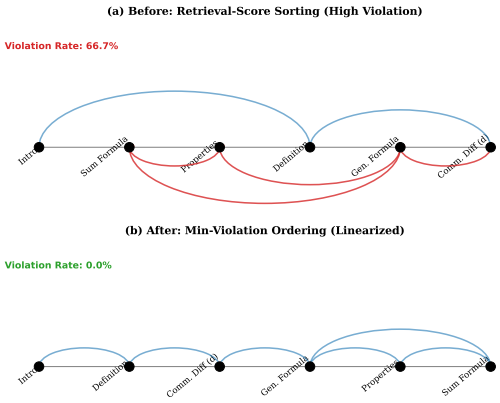

Figure 9: Visualization of the minimum-violation ordering effect. (a) **Before**: Sorting concepts solely by retrieval relevance scores results in a tangled dependency graph with frequent logical inversions. (b) **After**: SG-Planner's greedy ordering linearizes the pedagogical flow, reducing the violation rate for this segment from 50.0% to 0.0%, strictly adhering to epistemic precedence.

C.4 Comparison vs. Static Top-K Retrieval

Table 13 compares full SG-Planner against static Top-K retrieval. SG-Planner achieves higher recall (+4.6% @20) via expansion/pruning, and substantially lower violations (-3.7% total) via contextual filtering and ordering.

Method	Plan-Recall@K (%)			Violation Rate (%)		
	@10	@15	@20	Total	NEXT	BASED_ON
Static Top-K	68.4	74.2	79.1	12.1	7.2	4.9
SG-Planner	**72.1**	**78.5**	**83.7**	**8.4**	**4.1**	**4.3**

Table 13: Planner-level metrics (macro-average). Plan-Recall@K and precedence violation rates.

C.5 Correlation with Downstream Metrics & Failure Modes

On the validation set (**N=192 micro-segments from 48 videos**), Plan-Recall@20 positively correlates with KPC ($\rho = 0.62, p < 0.01$), while Violation Rate negatively correlates ($\rho = -0.48, p < 0.01$). This validates planner metrics as predictive proxies for pedagogical fidelity.

Failure Modes: SG-Planner fails on highly deictic transcripts (e.g., "from this...") yielding underspecified V_{sub}, or noisy graph edges forcing trade-offs between evidence and structure. Downstream KGVA can improve evidence grounding and ACP can better enforce the planned order, but neither

module can recover prerequisite nodes that are absent from the plan.

C.6 Oracle Planning Diagnostic

To quantify error propagation from Stage 1 and estimate the upper bound of downstream generation under ideal plans, we replace the predicted plan $P_i = (V_i^{sub}, E_i^{sub}, \pi_i)$ with an oracle plan while keeping Stage 2 (KGVA+ACP), Stage 3 (GKS), the backbone, frame budget, and decoding settings unchanged. For each phase segment, the oracle node set uses the gold knowledge-point concept set aligned to KnowledgeConcept nodes, the oracle edge set uses the induced prerequisite subgraph from the KnowledgeConcept layer, and the oracle order is obtained with the same minimum-violation ordering routine as SG-Planner. This oracle uses evaluation-only annotations as a diagnostic analysis only, and is not used for model selection or standard inference.

On the in-domain Test ID split, oracle planning improves KPC from 66.79 to 70.20, LOC from 63.11 to 66.10, and ROUGE-L from 36.62 to 37.05, corresponding to gains of +3.41, +2.99, and +0.43, respectively. This gap is consistent with the planner diagnostics in Appendix C.5, indicating that residual structural errors are still largely bottlenecked by Stage 1 planning quality.

D Implementation Details and Qualitative Evidence

This appendix addresses two reviewer concerns: (i) whether KGVA meaningfully leverages visual evidence beyond text conditioning, and (ii) whether ACP's constraints harm linguistic fluency. We provide implementation details, attention visualizations, and controlled ablations.

D.1 KGVA Architecture

KGVA inserts a gated cross-attention layer into the visual encoder **(e.g., InternVL3 or VideoLLaMA2)** after frame-level pooling and before global fusion. The visual backbone remains frozen during training. Given the plan-derived concept sequence $C_\pi = \text{TextEnc}(\{\text{name}(v) | v \in \pi\})$, visual tokens are updated as:

$$X_v' = X_v + \alpha \cdot \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V,$$

where $Q = X_v W_Q$, $K = C_\pi W_K$, $V = C_\pi W_V$, and α is a zero-initialized gate for stable training. Concept tokens use **frozen BGE-M3 embeddings** (name+definition concatenation), matching the encoder used in the retrieval stage to ensure semantic alignment between planning and grounding spaces.

D.2 Visualizing Dynamic Grounding (KGVA)

Does KGVA truly follow the pedagogical plan? Figure 10 visualizes the **temporal attention dynamics** over three distinct timesteps of a math lesson.

- **Baseline (Top Row)**: Exhibits *Visual Bias* (fixating on the instructor's face in T1) and *Visual Inertia* (failing to shift focus in T2/T3), leading to hallucinated content.
- **Ours (Bottom Row)**: Guided by the SG-Planner, KGVA dynamically shifts its hard attention mask from the **Definition** (T1) to the **Formula** (T2) and finally to the **Example** (T3).

This strictly aligns the visual encoder with the epistemic progression, proving that the Plan acts as a cognitive control signal for the visual encoder.

How bounding boxes are computed. For each timestep, we aggregate the cross-attention weights from the concept tokens to visual patch tokens (averaged over heads and layers used by KGVA), reshape them into a 2D spatial map, min-max normalize, then threshold the top-p mass (we use p=15%) to obtain a binary mask; the cyan box is the axis-aligned bounding rectangle of the largest connected component in this mask.

D.3 Vision Sensitivity Ablation

To quantify the reliance on visual information, we conducted a masking ablation (Figure 11). Masking the blackboard region causes a sharp performance drop in KCVR (ΔKPC $= -12.4$), whereas the baseline is less affected (ΔKPC $= -8.1$). This indicates that our model genuinely "reads" the blackboard rather than relying solely on language priors.

D.4 ACP Implementation Details

ACP (Adaptive Constrained Pedagogical Decoding) operates on the SG-Planner path $\pi = (c_1, \ldots, c_{|\pi|})$ and maintains a concept pointer k indicating the current planned concept c_k. At each decoding step t, the token logits are modulated as

$$z_t' = z_t + \beta \cdot \mathbf{1}_{A(c_k)} - \gamma \cdot \mathbf{1}_{F(\pi_{>k})},$$

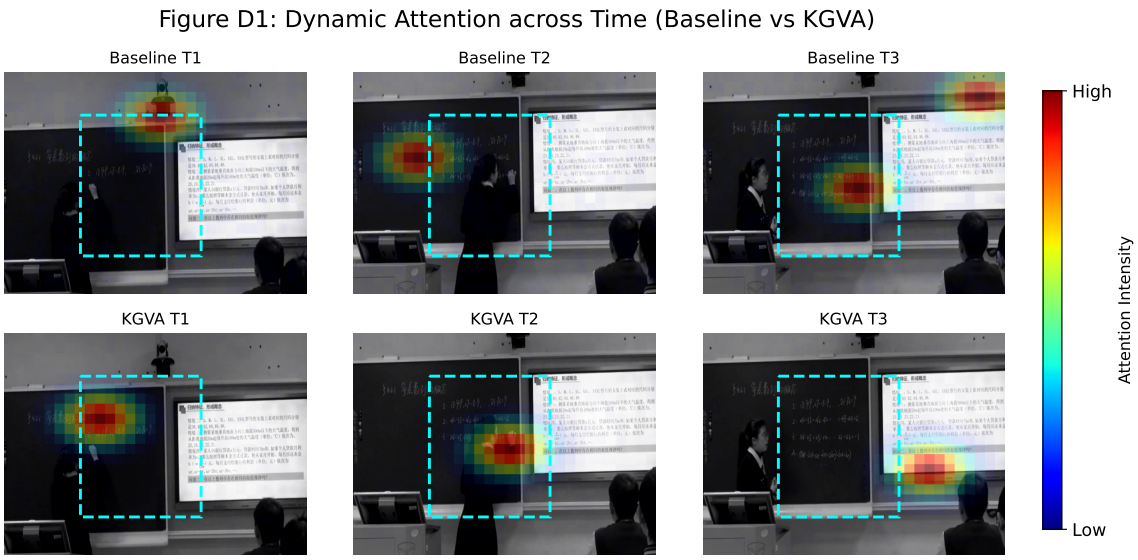

Figure 10: **Visualization of Plan-Guided Dynamic Grounding.** We compare the baseline (Top) vs. KGVA (Bottom) across three teaching phases. **Top Row:** The baseline suffers from visual bias (focusing on the face) and inertia. **Bottom Row:** KGVA creates attention-derived bounding boxes that track the pedagogical plan: (T1) Definition → (T2) Formula → (T3) Example. This pattern is consistent across our randomly sampled analysis set, where KGVA attention mass on blackboard regions is on average 2.4x higher than the baseline's attention.

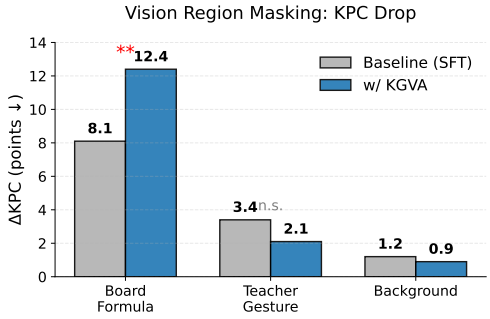

Figure 11: Vision sensitivity ablation **(on InternVL3 backbone)**. Masking board formulas hurts KPC significantly more for KCVR, confirming its strong dependency on visual evidence.

Table 14 shows that moderate constraint strength ($\gamma = 0.3$) achieves the best trade-off: KPC increases by +4.2 points over the Planner-only variant while ROUGE-L remains essentially unchanged.

Constraint γ	KPC	ΔKPC	R-L	ΔR-L
w/o ACP (0.0)	62.1	–	36.2	–
Weak (0.1)	64.5	+2.4	36.4	+0.2
Moderate (0.3)	**66.3**	**+4.2**	**36.1**	**-0.1**
Strong (0.5)	67.1	+5.0	35.4	-0.8

Table 14: Effect of constraint strength γ on in-domain validation. Moderate constraints deliver substantial structural gains (KPC) without degrading fluency (R-L).

where $A(c_k)$ denotes the alias set of surface forms associated with c_k, and $F(\pi_{>k})$ collects aliases for strictly future concepts $\{c_{k+1}, \dots\}$. The pointer advances to $k+1$ once the cumulative probability mass assigned to $A(c_k)$ exceeds a confidence threshold $\tau = 0.7$, effectively implementing a soft notion of "concept completion."

Triggering transition. Note that we do not force-insert discourse markers; instead, we monitor the generated tokens for a predefined set of transition keywords (e.g., 'Next', 'Then', 'Therefore') that naturally emerge from the LLM's language model, using them as soft triggers to update the pointer state alongside the probability threshold.

Per-module attribution. To isolate the respective contributions of KGVA and ACP under the same Planner path, we further evaluate three controlled variants on the in-domain validation set (48 videos, 192 phase segments), keeping all settings fixed except the Stage 2 module choice. KGVA provides plan-guided visual grounding, while ACP enforces topology-consistent realization by suppressing aliases of strictly future concepts during decoding. Table 15 shows that KGVA-only improves content fidelity and yields a modest KPC gain, whereas ACP-only delivers a larger structural gain by more directly reducing premature look-ahead. The combined setting performs best, indicating complementary rather than redundant

effects.

Variant	R-1	C-F1	KPC
Planner only	42.60	68.40	62.10
KGVA only	42.96	69.05	63.22
ACP only	42.85	69.20	64.43
KGVA+ACP	**43.30**	**69.80**	**65.60**

Table 15: Per-module attribution with the Planner fixed on the in-domain validation set. KGVA-only improves R-1/C-F1 and raises KPC by +1.12 over Planner only, while ACP-only yields a larger KPC gain of +2.33. Combining both gives the best overall Stage 2 result.

D.5 Qualitative Case Study: Logical Correction

To illustrate how KCVR corrects *Logical Inversion* in practice, Table 16 contrasts a representative micro-segment from the *Arithmetic Sequences* lesson. Both models have access to the same retrieval set and recover the correct terms, but they differ in how these pieces are ordered.

E Metric Definitions and Validation

This appendix provides formal definitions for our structure-aware evaluation metrics, along with implementation details and validation protocols to ensure reproducibility and alignment with human judgment (see E.2).

E.1 Metric Definitions

Knowledge Progression Consistency (KPC) KPC quantifies topological fidelity by measuring prerequisite violations in the generated summary S. Let $C_S = \{c_1, c_2, \ldots, c_m\}$ be the ordered sequence of concepts extracted from S via the Concept Trigger Layer (CTL). Let $G = (V, E)$ denote the ground-truth epistemic graph.

A violation occurs when a concept c_j precedes c_i despite $(c_i, c_j) \in E^+$, where E^+ is the transitive closure of prerequisite edges (primarily BASED_ON and PREREQ_OF). KPC is defined as:

$$\text{KPC}(S) = 1 - \frac{|V_{\text{err}}(S)|}{|P(C_S) \cap E^+| + \epsilon}, \quad (6)$$

where $V_{\text{err}}(S) = \{(c_j, c_i) \mid i < j \wedge (c_i, c_j) \in E^+\}$ and $P(C_S)$ is the set of all ordered pairs in C_S.

Edge Cases and Micro-Averaging. If $|P(C_S) \cap E^+| = 0$ (no comparable prerequisite pairs), the sample is marked *non-evaluable* for macro-averaging. For stable aggregation across the test set, we report the micro-averaged form:

$$\text{KPC} = 1 - \frac{\sum_S |V_{\text{err}}(S)|}{\sum_S |P(C_S) \cap E^+| + \epsilon}. \quad (7)$$

The $\epsilon = 10^{-4}$ term prevents division-by-zero. This micro-averaged formulation is robust to sparse graphs, ensuring that segments with fewer prerequisites do not disproportionately skew the metric (unlike macro-averaging). All KPC results in the paper use this formulation.

Learning Objective Coverage (LOC) LOC evaluates semantic alignment with ground-truth learning objectives $\mathcal{O}_{gt}$ across multilingual subjects. For each objective $o_i \in \mathcal{O}_{gt}$, we compute:

$$\text{LOC} = \frac{1}{|\mathcal{O}_{gt}|} \sum_i \mathbb{I}\left[\max_{s \in S_{\text{sent}}} \text{NLI}_{\text{entail}}(s, o_i) > \tau\right], \quad (8)$$

where $\text{NLI}_{\text{entail}}$ is the entailment score from a multilingual model (mDeBERTa-v3-base-mnli-xnli). The threshold $\tau = 0.7$ is calibrated on the validation set and held fixed across all subjects.

E.2 Human Evaluation Protocol Details

To ensure the rigor and reproducibility of the correlation results reported in Section 4.2, we implemented a strict **Double-Blind Evaluation Protocol**.

Annotator Demographics & Qualification. We recruited three independent evaluators, all of whom are graduate-level researchers with a minimum of two years of teaching assistant experience in STEM or Humanities disciplines. To mitigate potential bias, all annotators were strictly blinded to the model sources (Ours vs. Baselines vs. GPT-4o) and the specific research hypotheses. **All evaluators were compensated at a rate commensurate with standard graduate research assistant stipends at the host institution, which exceeds the local minimum wage.**

Sampling Stratification. The evaluation set comprises 50 randomly sampled videos, stratified to guarantee coverage across both domain and complexity dimensions. Specifically, we balanced the dataset between **In-Domain** (Math/Physics) and **Out-of-Domain** (History/InfoTech) subjects (25 videos each). Furthermore, to rigorously test the model's visual grounding capabilities, we stratified samples by **Visual Information Density**, selecting half from "High-Entropy" segments character-

Baseline: Text-RAG (High Retrieval, Low Logic)	Ours: KCVR (Planner-Constrained)
[Logical Inversion] The general term formula is $a_n = a_1 + (n-1)d$. ... Then the teacher explains that d **is called the common difference**...	[Correct Flow] The teacher first introduces the **definition** of an arithmetic sequence and the **common difference** d. Based on this, the teacher derives the general term formula $a_n = a_1 + (n-1)d$.
Critique: Presents the formula before defining its variables, violating the prerequisite chain.	*Critique:* Follows the SG-Planner path Definition $\rightarrow$ d $\rightarrow$ Formula, respecting epistemic precedence.

Table 16: Case study on an arithmetic-sequence segment. KCVR corrects the logical inversion observed in Text-RAG by enforcing the planner-derived concept order, yielding a definition→component→formula progression that aligns with expert pedagogy.

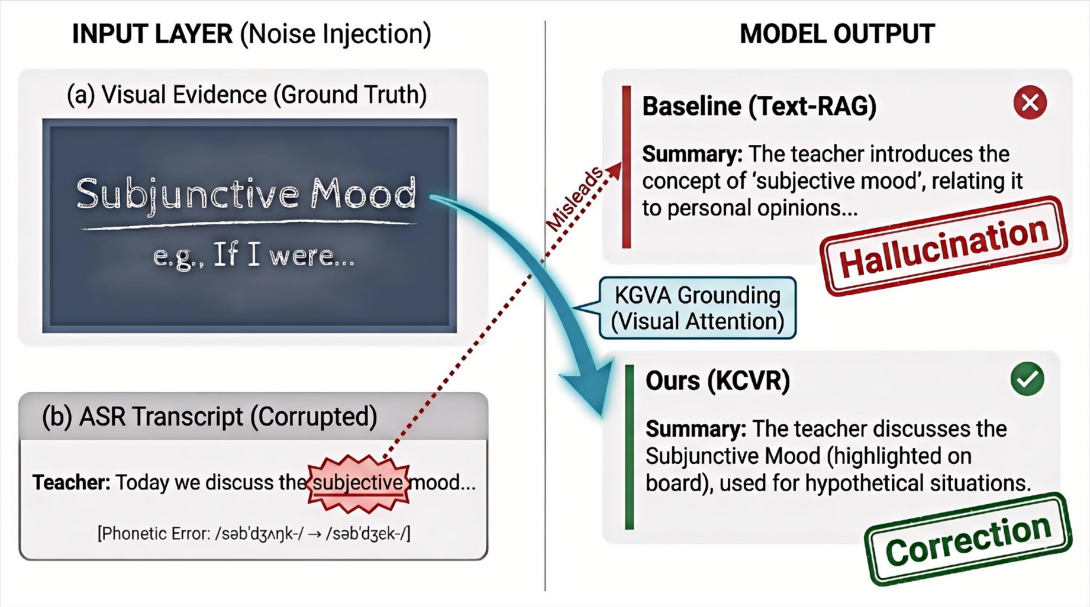

Figure 12: **Qualitative Analysis of Visual Grounding.** In this English Grammar case, the baseline hallucinates "Subjective" due to ASR corruption. In contrast, KCVR anchors to blackboard evidence ("Subjunctive") to rectify the error, validating the robustness gains reported in Table 5.

ized by dense blackboard derivations and half from "Low-Entropy" talking-head scenarios.

Evaluation Criteria. Annotators assessed each summary on a 5-point Likert scale across two primary dimensions. **Pedagogical Coherence** measures the logical validity of the narrative flow. A score of 5 (Perfect) indicates a rigorous prerequisite-consequence chain matching expert pedagogy; a score of 3 (Acceptable) denotes minor reordering that does not disrupt understanding; while a score of 1 (Severe Inversion) is assigned when conclusions prematurely precede their premises, causing significant cognitive dissonance. **Goal Fulfillment** evaluates the extent to which the summary addresses the ground-truth learning objectives. A score of 5 (Full Coverage) implies all key objectives are explicitly synthesized, whereas a score of 1 (Missed) indicates a failure to capture the primary pedagogical intent of the segment.

Agreement Analysis. To validate annotation reliability, we computed Fleiss' Kappa (κ) on the sampled subset. The agreement scores were $\kappa = 0.68$ for Pedagogical Coherence and $\kappa = 0.74$ for Goal Fulfillment, indicating substantial agreement among the expert evaluators. Disagreements in the final reported score were resolved by majority vote.

E.3 Robustness Case Study: The "Subjunctive" Error

While Appendix A.6 details our noise modeling, Figure 12 presents a concrete English grammar example. The ASR system misinterprets the phonetically similar keyword "*Subjunctive*" as "*Subjective*" (a common domain-shift error). **Text-RAG**, relying solely on the corrupt transcript, hallucinates a summary about "Subjective Opinions." **KCVR**,

Parameter	Value
Architectures (Full Model)	
Primary backbone	InternVL3-8B (frozen visual tower; 448px inputs)
Secondary backbone	VideoLLaMA2-7B (frozen visual tower; used for ablation)
Trainable components	KGVA modules + language-side adapters (LoRA r=32, α=64)
Text encoder	BGE-m3 (frozen; for concept embedding)
Architectures (Baselines)	
Transcript-only LLM	Qwen2.5-Instruct (used only for Qwen2.5-Instruct$_{\text{transcript}}$ baseline)
SG-Planner (Algorithm 1)	
Seed retrieval	Top-K concepts by s_{kc}, $K = 20$
Expansion hops	$h = 2$
Pruning budget	$B = 20$ nodes
Curriculum bias λ	0.5 (Selected on Val (ID); frozen for Test)
Violation weights w_r	[1.0, 5.0] for [BASED_ON, PREREQ_OF] (Selected on Val (ID))
Training	
Global Batch Size	128
Learning Rate	$2e^{-5}$ (Cosine decay, 3% warmup)
Optimizer	AdamW (DeepSpeed ZeRO-3)
Epochs	3 (Micro-SFT), 5 (Macro-SFT)
Inference	
Frames	8 frames/segment (Uniform Sampling)
OCR Channel	**Disabled** (Main Setting); Enabled only for Appendix Analysis
Decoding	Beam Search (beam=3), $\gamma_{ACP} = 0.3$, $\tau_{ACP} = 0.7$

Table 17: Hyperparameters for training and inference. This configuration is shared across both InternVL3 and VideoLLaMA2 backbones to ensure a fair comparison of structural reasoning capabilities.

guided by the SG-Planner and KGVA, grounds the query to the blackboard text "Subjunctive Mood" (visual evidence), demonstrating KGVA's capability to anchor generation to visual evidence when audio is ambiguous. While not immune to all noise, this mechanism explains the superior stability observed in Table 5.

E.4 Hyperparameters & Reproducibility

Table 17 lists the detailed configuration used to produce the main results. InternVL3-8B serves as the primary multimodal backbone. **Note that for the VideoLLaMA2 ablation reported in Table 1, we employed the identical hyperparameter set (including LoRA rank, learning rate, and planner budget) to verify framework robustness without model-specific tuning.** All planner hyperparameters are fixed before evaluation and are not tuned on the test set.

F Prompt Templates

This appendix presents the specific instructions used to align the Large Language Model with our neuro-symbolic modules. We employ a "Constraint-Injection" strategy, ensuring that the symbolic logic from the Knowledge Graph explicitly governs the neural generation process.

SG-Planner: Topological Constraint Injection. The SG-Planner instruction (Box F.1) is designed to penalize *Logical Inversion*. Crucially, we do not merely ask the model to sort concepts; we inject the sub-graph's dependency edges as explicit rules. The prompt mandates a "Constraint Citation" mechanism, requiring the model to explicitly reference which prerequisite rule dictates the current ordering. *Implementation Note:* The `{OCR_TOKENS}` field is populated only in the *OCR-Analysis* setting (see Appendix E.4); in the main setting, this field is hidden to enforce reliance on visual grounding.

Objectives-Aware Generator: Visual-Verbal Alignment To address *Visual Aphasia*, this

BOX F.1: SG-PLANNER INSTRUCTION (STAGE 1)

System Role: You are a Pedagogical Logic Verifier. Your task is to filter and linearize a bag of retrieved concepts into a strictly valid dependency chain.

Symbolic Context (Graph Injection):
- **Anchor Scope:** {TP_ANCHOR} (e.g., "Arithmetic Sequence Definition")
- **Candidate Pool:** {KC_CANDIDATES} (List of retrieved nodes with relevance scores)
- **Hard Constraints (Prerequisite Rules):**
{GRAPH_CONSTRAINTS} # e.g., "Concept A MUST precede Concept B"

Optimization Goals:
1. **Topological Validity:** You MUST NOT output a concept if its prerequisite (defined in Hard Constraints) has not appeared in a previous step.
2. **Evidence-Based Pruning:** Discard concepts that are topologically valid but lack explicit support in {TRANSCRIPT_SNIPPET} **[Analysis-Only: or {OCR_TOKENS}]**.
3. **Minimality:** Preserve only the minimal chain necessary to explain the Anchor Scope.

Output Format (JSON):
Return a JSON list. For each step, cite the constraint ID that justifies its position.
[{"step": 1, "concept": "...", "constraint_check": "Satisfies Rule A"}]

BOX F.2: OBJECTIVES-AWARE GENERATOR (STAGE 2)

Conditioning Context:
- **Pedagogical Goal:** {PEDAGOGICAL_GOAL} **[Note: Gold during Training; Self-Generated/Null during Inference]**
- **Focus Concept:** {PLANNED_NODE} (from Stage 1)
- **Visual Context:** (InternVL Visual Tokens projected by KGVA)

Generation Directives:
1. **Deictic Resolution (Critical):** The transcript contains ambiguous terms like "look at this". You MUST resolve these by describing the specific visual content currently on the blackboard.
2. **Objective-First Structure:** Start with a sentence explicitly linking the current visual state to the {PEDAGOGICAL_GOAL}.
3. **Visual Verification:** If the teacher mentions a concept but it is NOT written on the board, mark it as [Spoken-Only].
4. **Scientific Precision:** Transcribe all visible formulas in standard LaTeX.

Output: A micro-summary strictly grounded in the visual evidence.

Figure 13: **Constraint-Injection Prompts.** Box F.1: SG-Planner uses dynamic graph constraints to linearize prerequisites. Box F.2: The Generator resolves visual deixis conditioned on the planner's output. *Note on Leakage Control:* Fields marked *[Analysis-Only]* or *[Gold during Training]* are strictly withheld or replaced by model predictions during standard inference.

prompt (Box F.2) implements a "Deictic Resolution Protocol." It explicitly instructs the model to utilize the attention-guided visual features (via KGVA) to resolve ambiguous spoken terms (e.g., "this", "here") into concrete visual descriptions. *Leakage Control:* The {PEDAGOGICAL_GOAL} is supplied from ground-truth annotations during SFT (Training) to enable objective-aware supervision. At Inference, this field is either omitted or populated by a self-generated goal (via Chain-of-Thought), ensuring no test-time leakage.

V-Only Mode: Forensic Extraction In the transcript-free setting (Robustness Analysis), the visual encoder is repurposed to generate "Search Queries" for the Knowledge Graph (Box F.3). The prompt shifts focus from general captioning to *Knowledge Artifact Extraction*, enabling the SG-Planner to retrieve relevant subgraphs even without audio.

BOX F.3: VISUAL FORENSIC FOR QUERY GENERATION (V-ONLY)

Role: You are a Visual Forensic Analyst converting classroom frames into Knowledge Graph search queries.

Instruction:
Analyze the sequence of 8 frames. Ignore the teacher's appearance. Focus exclusively on the **Knowledge Artifacts** (Text, Diagrams, Formulas) on the blackboard/slides.

Extraction Protocol:
1. **Handwriting Transcription:** Transcribe all legible keywords.
2. **Structure Recognition:** Is it a Definition? A Proof?
3. **Layout Logic:** Describe the spatial flow.

Output (Search Query Format):
Query Keywords: [List of transcribed terms for KG lookup]
Pedagogical Intent: [Inferred teaching phase]

Figure 14: **V-Only Robustness Prompt.** Used to generate surrogate text queries for SG-Planner when transcripts are withheld.