

MOSAIC: A Large-Scale Multimodal Open-Course Segmentation and Alignment Corpus in Chinese

Yuming Ai¹

Shuai Fan²

Hua Xu²

Fang Kong^{1,3*}

¹Soochow University

²AISpeech Ltd.

³Jiangsu Key Lab of Language Computing, Suzhou 215123, China

20244227037@stu.suda.edu.cn, shuai.fan@aispeech.com

hua.xu@aispeech.com, kongfang@suda.edu.cn

Abstract

Real-world lectures follow a predefined topic inventory (e.g., a syllabus), but instructors often reorder, paraphrase, or skip items, creating mismatches between the outline and the spoken content. Lectures also contain non-topical utterances (openings, transitions and wrap-ups) that should not be assigned to any topic, yet existing corpora rarely model both challenges. We present **MOSAIC** (Multimodal Open-course Segmentation and Alignment corpus In Chinese), a large-scale Chinese MOOC corpus for *topic-guided segmentation and alignment* under predefined inventories, with explicit *non-topical* labels and matched slide resources. MOSAIC includes **MOSAIC-G** (human-annotated gold benchmark with utterance-level topic alignment and video-level links to matched slide resources) and **MOSAIC-S** (automatically labeled silver data), collectively covering 179 courses and 14,942 videos. To scale annotation, we propose **DORA** (Dual-stage Outline Refinement and Alignment), a two-stage pipeline that refines noisy inventories and then performs joint segmentation and topic assignment. Experiments show that silver supervision substantially improves the performance of local models and that incorporating slides consistently benefits knowledge-point extraction, a lecture-comprehension proxy.¹

1 Introduction

Long instructional videos are hard to follow without structure, which is why most real-world lectures come with a *predefined topic inventory*—an outline, syllabus, or knowledge-point list that constrains what the lecture covers and supports downstream tasks such as navigation, retrieval, and summarization. However, an outline is not a script: instructors

*Corresponding author.

¹Code and release information for the public annotations and metadata are available at <https://github.com/AmyIvan/MOSAIC>.

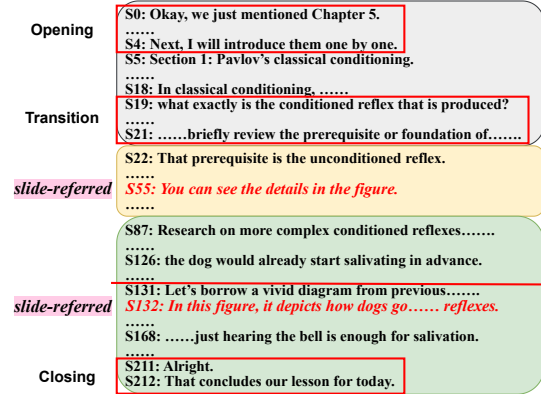


Figure 1: Without predefined-topic inventory constraints, confusable topics mix and non-topical utterances (e.g., opening, transition, and closing) are absorbed, causing shifted or missed boundaries (red line: missed boundary); slide-referred cues (e.g., “see the figure”) further highlight matched slide resources as useful context for lecture understanding. See Fig. 2 for correct segmentation under predefined-inventory refinement.

routinely reorder, merge, paraphrase, or skip items during spontaneous delivery, causing mismatches between the predefined list and the actual lecture content. This creates a *noisy-inventory* setting where accurate topic assignment becomes as critical as boundary placement. Meanwhile, lectures are not continuous topical exposition: openings, transitions, and wrap-ups are indispensable for discourse organization, yet they do not belong to any predefined topic and must be explicitly modeled.

Most segmentation benchmarks do not capture this setting. Many assume topic-free segmentation or ignore fixed inventories during evaluation, leaving common lecture failure modes under-measured. Figure 1 illustrates a typical breakdown²: without topic-guided refinement, systems conflate similar topics and merge segments that should map to different topics; they also absorb non-topical

²For readability, we present dataset-related fields and examples in English throughout the paper.

utterances into adjacent topics, shifting true boundaries and corrupting assignments. Moreover, slides present diagrams, formulas, and visual demonstrations difficult or impossible to convey through speech alone—instructors frequently abbreviate or skip such content, making matched slide resources important for complete lecture understanding.

To fill this gap, we introduce **MOSAIC** (**M**ultimodal **O**pen-course **S**egmentation and **A**lignment corpus **I**n **C**hinese), a large-scale Chinese MOOC corpus for topic-guided segmentation under predefined inventories (see Table 2 for comparison). MOSAIC comprises **MOSAIC-G** (gold, human-annotated benchmark) and **MOSAIC-S** (silver, automatically labeled training data). Both provide utterance-level labels for joint segmentation and topic *alignment* (i.e., assigning each segment a predefined topic or non-topical status), supporting rigorous evaluation and large-scale training. MOSAIC also records video-level links to matched slide resources for multimodal grounding: all MOSAIC-G videos are associated with corresponding slide resource(s), while MOSAIC-S provides automatic video-level links where available.

To make annotation scalable, we propose **DORA** (**D**ual-stage **O**utline **R**efinement and **A**lignment), a two-stage pipeline that first refines potentially mismatched inventories conservatively, then performs joint segmentation and alignment. DORA enables large-scale silver labeling beyond the gold benchmark. Experiments demonstrate that silver supervision substantially improves local models, and that incorporating slides consistently benefits knowledge-point extraction, a lecture-comprehension proxy. Our main contributions are:

- **A benchmark for topic-guided segmentation with explicit non-topical modeling.** We introduce MOSAIC, a large-scale Chinese MOOC corpus with utterance-level labels for joint boundary detection and topic alignment under predefined inventories. MOSAIC explicitly annotates non-topical lecture utterances (e.g., openings, transitions, and wrap-ups), recognizing them as integral components of natural lecture delivery rather than noise.
- **Multimodal slide resources for grounded lecture understanding.** MOSAIC records video-level links to matched slide resources that reflect how learners follow lectures with both speech and visuals. Experiments

on knowledge-point extraction—a lecture-comprehension proxy—show that integrating slide text substantially improves the quality of extracted knowledge points.

- **A scalable annotation pipeline with validated output quality.** We introduce DORA, a two-stage pipeline that first refines potentially mismatched topic inventories, then performs joint segmentation and alignment. Experiments demonstrate that a 0.6B model fine-tuned on MOSAIC-S labels outperforms its 32B counterpart, approaching strong closed-source LLMs with orders of magnitude fewer parameters, a substantial efficiency gain.

2 Related Work

Topic Segmentation. Classical segmentation approaches rely on lexical cohesion or probabilistic topic models (Hearst, 1997; Choi, 2000; Utiyama and Isahara, 2001; Eisenstein and Barzilay, 2008). Neural methods often formulate segmentation as boundary prediction or sequence labeling (Koshorek et al., 2018; Arnold et al., 2019), which becomes more challenging in spoken transcripts due to automatic speech recognition (ASR) noise (Hsueh et al., 2006). Recent work explores transcript segmentation for structured navigation in long-form spoken content (Freisinger et al., 2025). However, most existing benchmarks focus on topic-free segmentation or do not require mapping segments to a predefined inventory, leaving the practical setting of *topic-guided* segmentation under-explored. Beyond linear segmentation, another line of work recovers *hierarchical* topical structure, from multi-scale lexical cohesion (Eisenstein, 2009) and affinity-propagation-based hierarchical segmentation (Kazantseva and Szpakowicz, 2014) to section-level neural models (Arnold et al., 2019) and table-of-contents generation over spoken transcripts (Freisinger et al., 2025). MOSAIC is complementary to this line: the benchmark task is deliberately flat—the inventory is given per video, and the goal is faithful joint segmentation and alignment to it, including rejecting uncovered candidates and isolating non-topical spans—while hierarchy lives one level up in the corpus (course→chapter→video→topic organization plus course-level knowledge graphs), enabling future hierarchical, inventory-guided modeling.

Instructional corpora. Large instructional corpora provide narration or weak temporal supervision

(Sanabria et al., 2018; Miech et al., 2019), while procedural datasets (e.g., YouCook2/COIN/CrossTask) emphasize step boundaries rather than outline-grounded discourse (Zhou et al., 2018; Tang et al., 2019; Zhukov et al., 2019). Educational resources such as MOOCCube/MOOCCubeX and Lecture-Bank support prerequisite learning or slide-based materials (Yu et al., 2020, 2021; Li et al., 2019), but typically lack utterance-level topic alignment under predefined inventories. Meanwhile, recent work shows that multimodal cues can improve topic segmentation in videos (Yu et al., 2025), and long-form multimodal meeting corpora have also been released to support realistic transcription (Chen et al., 2025). Despite these advances, existing instructional corpora lack the combination of predefined topic inventories, utterance-level topic alignment, explicit non-topical modeling, and matched slide resources that characterizes real lecture consumption. Closest to our setting, Problem-Oriented Segmentation and Retrieval (POSR; Wang et al., 2024) performs candidate-worksheet-problem-guided segmentation and problem identification over English tutoring transcripts, which we use for external validation (Table 8). In other languages, the PASTEL corpus (Mdhaaffar et al., 2020) offers ten hours of French lectures with manual transcripts, thematic segmentation, and slide materials, at a much smaller scale than MOSAIC and without predefined topic inventories or utterance-to-inventory alignment.

3 Construction of MOSAIC

3.1 Data Source

MOSAIC is constructed from public courses on iCourse163³, a major Chinese MOOC platform popular among millions of students and instructors. We focus on National First-Class and Exemplary Courses to ensure high instructional quality and well-formed supporting materials for annotation.

The raw data follows a four-level hierarchy with metadata: a **Course** carries prior-knowledge requirements, objectives, and a concept-prerequisite knowledge graph; a **Chapter** groups videos and their shared slide decks (PDF/PPT); a **Video** (lesson) provides the ASR transcript (timestamped utterances), an instructor-provided knowledge-point outline—the predefined topic inventory—and a summary; a **Topic** is a knowledge point whose utterance alignment is the corpus’s central annotation target.

³<https://www.icourse163.org/>

Importantly, while the platform provides these materials, **no high-quality alignment exists between them**. We construct all utterance-to-topic alignments from scratch, and record links to matched slide resources at video level: all MOSAIC-G videos are associated with corresponding slide deck(s), while MOSAIC-S provides automatic video-to-slides links where available. Table 1 summarizes the corpus scale; the corpus spans 10 academic categories—Sciences (34.1%), Engineering (32.4%), Economics (10.6%), and seven other categories—supporting evaluation across diverse pedagogical styles and domain vocabularies.

Metric	Value
<i>Scale</i>	
courses	179
chapters	1,896
videos	14,942
duration (hours)	7,354.93
Avg duration / video (min)	29.53
Avg videos / chapter	7.88
Avg chapters / course	10.59
<i>Transcripts</i>	
Avg utterances / video	186.93
<i>Topic Outlines</i>	
Avg topics / video	4.30
<i>Slides</i>	
Videos with slide decks	10,385 (69.5%)
matched slide decks	7,976
slide deck pages	175,951
Avg pages / video	16.94

Table 1: Raw corpus statistics for MOSAIC.

3.2 Target Task Definition

We define MOSAIC’s topic-guided segmentation task. Given a sequence of timestamped transcript utterances $\{u_i\}_{i=1}^N$ and a candidate topic inventory $\mathcal{T} = \{T_k\}_{k=0}^{K-1}$, the goal is to: (i) partition the utterances into contiguous segments, and (ii) assign each segment either a topic index from $\{0, \dots, K-1\}$ or a dedicated non-topical label for utterances (e.g., openings, transitions, wrap-ups) that do not belong to any predefined topic.

We operationalize the output as per-utterance labels $\mathbf{w} = [w_1, \dots, w_N]$ with $w_i \in \{-1, 0, 1, \dots, K-1\}$, where -1 denotes non-topical utterances (e.g., openings, transitions and wrap-ups). This labeling encodes boundaries and utterance-to-topic

alignment, enabling evaluation of segmentation quality and topic assignment accuracy.

This formulation captures two realistic challenges: inventories may not perfectly match the realized lecture, creating semantically similar candidates that complicate assignment; and non-topical utterances form a substantial portion of real lectures (4–5% of utterances; Appendix D) that must be modeled explicitly rather than forced into adjacent topics.

Other potential tasks. MOSAIC’s annotations also support several related tasks: lecture-script generation, multimodal grounding, curriculum-aware modeling, and prerequisite-relation mining via the course-level knowledge graphs; we leave these extensions, including fine-grained utterance-to-slide-page grounding, to future work.

3.3 DORA: Automatic Annotation Pipeline

The large scale of MOSAIC (179 courses, 14,942 videos) makes fully manual annotation impractical. We therefore propose **DORA** (Figure 2), a two-stage automatic annotation pipeline that produces high-quality labels at scale: Stage 1 refines the candidate inventory to reduce mismatches between the predefined list and the realized lecture, and Stage 2 performs joint topic segmentation and alignment under the refined inventory.

3.3.1 Stage 1: Topic List Refinement

Implementation. Given transcript utterances and the original topic list $\mathcal{T}$, DORA outputs a refined, time-ordered topic list $\mathcal{T}'$ for Stage 2 decoding. We implement Stage 1 using large language models (LLMs) via structured prompting. The refinement proceeds in two sequential steps. First, **conservative reconstruction**: the LLM filters topics with no substantive support in the transcript, deduplicates near-duplicates by retaining one representative candidate, and reorders only when temporal evidence is clear, following a recall-first principle. Second, **reflection audit**: the LLM re-examines the filtered list against the transcript as the sole evidence source, restoring any mistakenly removed topics and removing spurious keeps or unsupported edits. This two-step design leverages the LLM’s ability to reason about semantic similarity and temporal ordering while mitigating over-aggressive pruning through explicit self-correction. Detailed prompt templates are provided in Appendix A.

3.3.2 Stage 2: Segmentation and Alignment

Implementation. Given the transcript and the refined list $\mathcal{T}'$, Stage 2 outputs the per-utterance labels w over $|\mathcal{T}'|$ topics, using the Stage 1 backbone with a structured prompt (Appendix A) that enforces near-monotonic topic progression and explicit non-topical labeling; the output JSON segment triplets (start, end, topic_id) are deterministically converted to per-utterance labels for evaluation.

3.4 Annotation Schema

Utterance-to-topic alignment with explicit non-topical utterances. Each utterance receives a topic index or the dedicated non-topical label (−1; openings, transitions, wrap-ups), preventing contamination of topical segments and boundary drift.

Slide-resource grounding. Lecture understanding is inherently multimodal: learners follow speech together with slides. MOSAIC therefore records video-level links to matched slide resource(s)—manually verified for all MOSAIC-G videos and linked automatically for MOSAIC-S—while slide-derived text is used where available and permitted for slide-aware modeling.

Table 2 situates MOSAIC among representative corpora. While large-scale resources such as HowTo100M provide weak supervision and procedural datasets annotate step boundaries, MOSAIC uniquely combines predefined topic inventories, utterance-level joint segmentation, explicit non-topical utterances, and matched slide resources.

3.5 Annotation Process and Quality

We construct MOSAIC in two phases: manual gold annotation (MOSAIC-G), then automatic DORA annotation of remaining courses (MOSAIC-S).

MOSAIC-G: Gold annotations. We select 6 courses across psychology, medicine, biology, economics, arts, and engineering for fully manual annotation from scratch. Selection prioritized disciplinary diversity, clear topic structure, moderate length (10–20 min), and reliable ASR quality (see Appendix F for the operational protocol and measured quality proxy). We refer to these 6 gold courses as IT, PP, PA, PN, EP, and MM, whose full names are given in Table 3. Two annotators independently label each video without automatic pre-annotation, assigning utterance-level topic labels and marking non-topical segments. Three reviewers adjudicate disagreements jointly to produce final gold labels. For the slide modality, all 234 MOSAIC-G videos

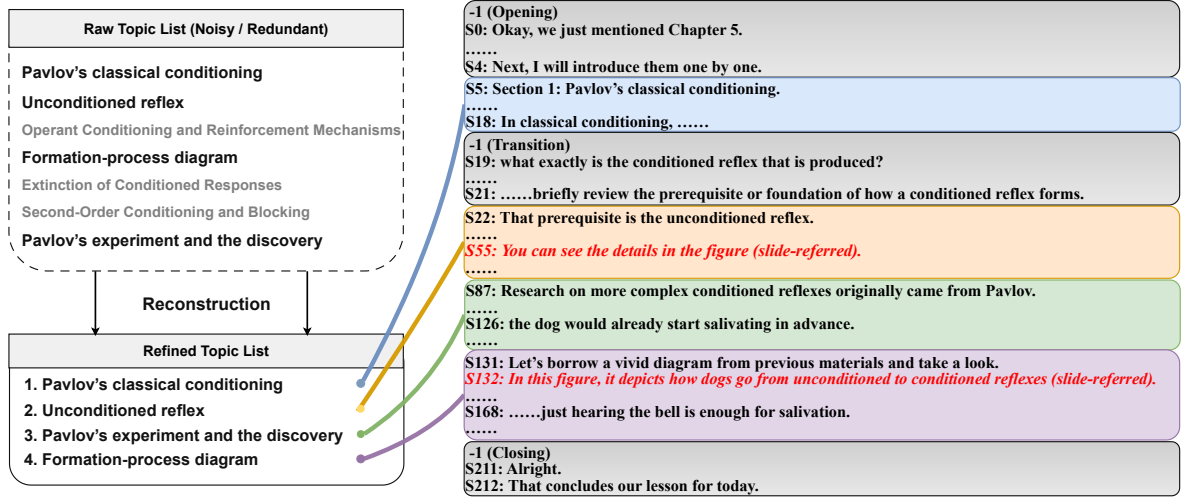


Figure 2: DORA overview: Stage 1 refines the predefined topic inventory (conservative reconstruction + reflection audit) to filter confusable distractors; Stage 2 decodes per-utterance labels for joint segmentation and topic alignment, explicitly separating non-topical lecture utterances. See Appendix B for the complete annotated transcript.

Dataset	Lang.	Scale	Hier.	Transc.	Course-topic inventory	Text-Align	Slides	KG
MOSAIC (ours)	ZH	14,942 videos; 7,355h	✓	✓	✓	✓	✓	✓
LessonLink (Wang et al., 2024)	EN	405h	✗	✓	~	✗	✗	✗
PASTEL (Mdhaaffar et al., 2020)	FR	10h	~	✓	✗	✓	✓	✗
MOOCCube (Yu et al., 2020)	ZH	706 courses	~	✓	~	✓	✗	✓
MOOCCubeX (Yu et al., 2021)	ZH	4,216 courses	✓	✓	~	✓	✗	✓
LectureBank (Li et al., 2019)	EN	1,352 files	~	✗	~	✗	✓	✓
How2 (Sanabria et al., 2018)	EN+PT	2,000h	✗	✓	✗	✓	✗	✗
HowTo100M (Miech et al., 2019)	EN	136M clips	~	✓	✗	✓	✗	✗
YouCook2 (Zhou et al., 2018)	EN	2,000 videos	✗	✗	✗	✓	✗	✗
COIN (Tang et al., 2019)	EN	11,827 videos	✓	✗	~	✓	✗	✗

Table 2: Comparison with representative datasets. Hier.=Hierarchical structure, Transc.=Transcripts, Text-Align=fine-grained text/video segment alignment, Slides=publicly released matched slide resources or slide-deck links, KG=Knowledge graph. ✓=Yes, ~=Partial, ✗=No. LessonLink provides a per-session candidate set of SAT (Scholastic Assessment Test) worksheet problems, but not an instructor-authored hierarchical course-topic inventory.

Course	Domain	Videos	Hours	T	NT	Pages	Utt/V	Top/V
EP	Medicine	28	6.1	156	56	333	255	5.5
IT	Economics	22	7.1	118	70	431	269	5.2
MM	Engineering	30	5.9	169	81	318	225	5.6
PA	Arts	54	9.4	343	150	263	222	6.3
PN	Biology	48	7.1	256	90	1,250	176	5.3
PP	Psychology	52	12.9	260	138	239	250	4.9
Total	–	234	48.4	1,302	585	2,834	228	5.5

Table 3: MOSAIC-G statistics. EP=Epidemiology, IT=International Trade, MM=Mechanics of Materials, PA=Photographic Art, PN=Plant Nutrition, PP=Personality Psychology. T/NT=Per-video counts of contiguous annotated topical/non-topical segments, Pages=Matched-deck pages, Utt/V=Mean utterances, Top/V=Mean topics.

have verified video-level links to matched slide resource(s); no manual utterance-to-slide-page labels are added. All human annotators and adjudicators were members of our research laboratory rather than externally recruited workers; no crowdsourcing plat-

form was used, and no separate annotation-specific compensation was provided.

MOSAIC-S: Silver annotations. The remaining 173 courses (14,708 videos; 7,307 hours) are annotated fully automatically without any subsequent

Metric	MOSAIC-G	MOSAIC-S
<i>Scale</i>		
courses	6	173
chapters	48	1,848
videos	234	14,708
duration (hours)	48.41	7,306.52
<i>Segmentation Annotations</i>		
topical segments	1,302	58,431
non-topical segments	585	21,019
total segments	1,887	79,450
<i>Per-Video Averages</i>		
Avg duration (min)	12.41	29.81
Avg utterances	227.71	186.28
Avg topics	5.52	4.28
Avg segments	8.06	5.40
<i>Slides</i>		
Videos with slide decks	234 (100%)	10,151 (69%)
slide deck pages	2,834	173,117

Table 4: Annotation statistics for MOSAIC-G (gold) and MOSAIC-S (silver). Topical and non-topical segments are contiguous label runs: adjacent utterances with the same label form one segment. Slide pages denote pages in matched decks, not utterance-to-page labels. Non-topical utterances constitute a meaningful portion of real lectures: on MOSAIC-G, they account for 4.6% of utterances and 5.2% of total time (Appendix D).

manual correction. For utterance-to-topic alignment, we apply DORA to generate per-utterance labels; we validate the utility of these silver labels through downstream fine-tuning, where models trained on MOSAIC-S achieve strong performance on the gold benchmark (Section 4). For video-to-slide-deck linking (available for 69% of videos), we use an automatic pipeline combining title matching, rule-based filtering, and LLM verification.

Tables 3 and 4 report summary statistics for the MOSAIC-G and MOSAIC-S subsets. On average, a realized topic spans ≈ 41 utterances on MOSAIC-G and ≈ 44 on MOSAIC-S, and corresponds to ≈ 1.0 contiguous segment; the task difficulty therefore lies in discriminating semantically confusable candidates and placing boundaries precisely, rather than in topic fragmentation.

Inter-annotator agreement. To quantify label quality, we randomly sampled a subset of 24 videos, comprising 5,592 utterances in total, drawn from all 6 gold courses for our inter-annotator agreement evaluation. The same two annotators as in the gold annotation independently labeled these videos from

scratch. As reported in Table 5, utterance-level topic assignment reaches almost perfect agreement between the two annotators, with a Cohen’s κ of 0.87. Boundary agreement is lower, with a Boundary F1 of 0.60 under exact match and 0.75 under a more lenient ± 1 -utterance tolerance. Both annotators also agree strongly with the final gold labels ($\kappa = 0.85/0.81$), further confirming the reliability of our overall annotation process.

4 Experiments

4.1 Evaluation Metrics

Our joint task requires placing boundaries and mapping each utterance to the correct topic under a predefined inventory, while keeping non-topical utterances separate. For segmentation we adopt two standard metrics, P_k (Beeferman et al., 1999) and WinDiff (Pevzner and Hearst, 2002), both computed over a sliding window of size $k = \lfloor N/2R \rfloor$ (N utterances, R reference segments); lower is better. Segmentation alone is insufficient, since plausible boundaries may still carry wrong topic assignments under confusable candidates. We therefore adapt the Segmentation and Retrieval Score (SRS) (Wang et al., 2024) to our topic-alignment setting: its line-based form is instantiated as per-utterance topic-label accuracy, and Time-SRS is its duration-weighted counterpart; higher is better. All four metric definitions appear in Appendix H.

4.2 Experimental Settings

Evaluation modes for non-topical utterances.

Non-topical utterances, such as openings, transitions, and wrap-ups, are essential for natural lecture delivery, yet their exact boundary locations can be slightly ambiguous even for human annotators in practice, particularly near topic transitions. To reflect different tolerances, we evaluate non-topical labeling ($w_i = -1$) under two modes. **Hard** requires exact match: a prediction is correct only if it also assigns -1 to the utterance. **Soft** accepts predictions that label non-topical utterances as -1 or as an adjacent topic (the nearest topical neighbor on either side), better matching downstream use cases where small boundary shifts are acceptable. For topical utterances ($w_i \geq 0$), both modes require exact topic match. We report both modes throughout to characterize model behavior under both strict and more practical evaluation criteria.

Topic inventory settings. We evaluate under two complementary topic-inventory conditions in order

Metric	Value	Metric	Value	Metric	Value
Cohen’s κ (topic assignment)	0.87	Boundary F1 score (exact match)	0.60	Annotator 1 vs. Gold (κ)	0.85
Krippendorff’s α	0.87	Boundary F1 (± 1 tolerance)	0.75	Annotator 2 vs. Gold (κ)	0.81
Raw agreement (accuracy)	89.4%	Videos evaluated	24	Utterances evaluated	5,592

Table 5: Inter-annotator agreement on MOSAIC-G. Two annotators independently labeled 24 videos, achieving almost perfect agreement ($\kappa > 0.8$).

Course	SRS ($\uparrow$)				Time-SRS ($\uparrow$)				WinDiff ($\downarrow$)				P_k ($\downarrow$)			
	Clean		Noisy		Clean		Noisy		Clean		Noisy		Clean		Noisy	
	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard
IT	0.85	0.80	0.76	0.72	0.85	0.80	0.77	0.73	0.23	0.23	0.31	0.30	0.17	0.16	0.23	0.22
PP	0.78	0.76	0.76	0.75	0.78	0.76	0.76	0.75	0.25	0.24	0.27	0.27	0.18	0.18	0.20	0.19
PA	0.87	0.86	0.85	0.84	0.87	0.86	0.85	0.84	0.22	0.22	0.22	0.21	0.16	0.16	0.16	0.16
PN	0.87	0.86	0.83	0.83	0.86	0.86	0.83	0.83	0.22	0.21	0.24	0.23	0.18	0.17	0.19	0.18
EP	0.82	0.81	0.81	0.81	0.82	0.82	0.81	0.81	0.25	0.24	0.30	0.29	0.18	0.17	0.20	0.19
MM	0.79	0.78	0.75	0.74	0.79	0.78	0.75	0.75	0.27	0.26	0.27	0.26	0.22	0.21	0.21	0.20
Avg	0.83	0.81	0.79	0.78	0.83	0.81	0.80	0.79	0.24	0.23	0.27	0.26	0.18	0.18	0.20	0.19

Table 6: Stage-2-only results on MOSAIC-G with Qwen3-32B, using the original (Clean) or sibling-augmented (Noisy) inventory and Soft/Hard handling of non-topical utterances. Performance degrades under the sibling-augmented inventory (average Soft SRS: 0.83 $\rightarrow$ 0.79), with IT showing the largest drop (-9 percentage points) due to highly confusable economic terminology. $\uparrow$ = higher is better; $\downarrow$ = lower is better.

to assess the robustness of each method. **Clean** uses the correct predefined inventory exactly as provided by the course instructors, without any modification. **Noisy** augments the same inventory with additional semantically confusable candidates automatically generated by GPT-4o-mini, simulating imperfect real-world scenarios where topic lists may contain redundant, overlapping, or otherwise misaligned items. Specifically, for each video with K gold topics, we inject K plausible but not-covered distractors (see Appendix E for generation details), i.e., one hard negative per gold topic, keeping the overall difficulty comparable and balanced across videos of different lengths. This distinction is crucial because the task requires both accurate boundary detection *and* selecting the correct topic from among many competing candidates—noisy inventories therefore stress-test this latter capability more directly.

4.3 Baseline Results and Analysis

Table 6 reports per-course Stage-2-only performance across all six MOSAIC-G courses with a Qwen3-32B backbone under the original and sibling-augmented inventory conditions. Under the original inventory, Stage 2 achieves an average SRS of 0.83 (Soft) and 0.81 (Hard). With the sibling-augmented inventory, average Soft SRS drops from 0.83 to 0.79, and Soft WinDiff increases from 0.24

to 0.27, showing that confusable candidates degrade both topic assignment and boundary detection.

Cross-course difficulty analysis. Under the sibling-augmented inventory, IT drops the most (-9 percentage points (pp), reflecting its dense and highly confusable economics terminology), MM and PN drop moderately (~ 3 – 4 pp), while PA, PP, and EP remain comparatively stable across conditions (≤ 2 pp). The Soft–Hard SRS gap is small for five courses, but IT is a clear exception (about 5 pp under Clean and 4 pp under Noisy).

4.4 Ablation: Effect of Topic Refinement

To more clearly isolate the specific contribution of retained Stage 1 reconstruction, we compare Stage 1 without reflection against Stage 2 only.

Table 7 shows that adding Stage 1 reconstruction without reflection improves alignment and segmentation across most of the evaluated courses. When averaged over all 6 courses, SRS/Time-SRS increase while WinDiff/ P_k decrease, indicating that Stage 1 reconstruction mitigates candidate confusion in the LLM-doubled inventory before the final decoding stage begins. The improvement is particularly pronounced on IT and PA, while EP shows mixed results, suggesting that the benefit depends on the degree of inventory–lecture mismatch.

Course	Stage 2 only				Stage 1 w/o reflection			
	SRS ($\uparrow$)	Time-SRS ($\uparrow$)	WinDiff ($\downarrow$)	P_k ($\downarrow$)	SRS ($\uparrow$)	Time-SRS ($\uparrow$)	WinDiff ($\downarrow$)	P_k ($\downarrow$)
IT	0.772	0.766	0.299	0.240	0.811	0.811	0.255	0.213
PA	0.773	0.768	0.274	0.241	0.812	0.807	0.204	0.172
PP	0.755	0.755	0.270	0.195	0.790	0.791	0.290	0.200
PN	0.814	0.814	0.271	0.210	0.850	0.853	0.261	0.192
EP	0.784	0.786	0.301	0.206	0.782	0.784	0.271	0.200
MM	0.747	0.746	0.274	0.243	0.763	0.765	0.248	0.203
Avg	0.774	0.773	0.282	0.222	0.801	0.802	0.255	0.197

Table 7: Hard-evaluation ablation on MOSAIC-G with Qwen3-32B and the LLM-doubled inventory: Stage 2 only vs. Stage 1 without reflection across all six evaluated benchmark courses. Bold indicates the best value in each course and metric. This inventory and run differ from the sibling-augmented setting in Table 6.

4.5 External Validation: LessonLink Benchmark

We further evaluate on LessonLink (Wang et al., 2024) as an external testbed for candidate-worksheet-problem-guided segmentation and problem identification. As shown in Table 8, under the same backbone (GPT-4-turbo), DORA improves over POSR (Wang et al., 2024) on SRS/Time-SRS, indicating fewer candidate-confusion errors when segmentation-only scores are similar. With GPT-4o-mini, we obtain further gains, demonstrating both DORA’s generality beyond Chinese MOOCs to English tutoring and its scalability to different backbone models, on which the strongest backbone gives the best results in this comparison.

Metric	POSR GPT-4-turbo	DORA GPT-4-turbo	DORA GPT-4o-mini
SRS ($\uparrow$)	0.88	<u>0.90</u>	0.93
Time-SRS ($\uparrow$)	0.89	<u>0.91</u>	0.93
WinDiff ($\downarrow$)	0.32	0.32	0.25
P_k ($\downarrow$)	0.16	<u>0.14</u>	0.10

Table 8: Results on LessonLink dataset (Wang et al., 2024). POSR and DORA are methods; LessonLink is the evaluation corpus. Bold indicates the best value; underline indicates the second-best value.

4.6 Silver Label Quality Validation

We assess whether DORA-generated silver labels provide meaningful supervision for downstream model training rather than superficial artifacts that exploit shallow statistical patterns. As a reference point, classical pipelines (TextTiling for segmentation followed by BM25/ColBERT retrieval) perform poorly (SRS < 0.15) due to structural mismatch with joint segmentation–alignment (Appendix C).

Fine-tuning with silver supervision. We fine-tune Qwen3-0.6B with Low-Rank Adaptation (LoRA) (Hu et al., 2022) on MOSAIC-S (14,708 videos) and evaluate on MOSAIC-G. To rule out trivial statistical gains, we construct a shuffled control permuting utterance-to-topic assignments within each video while preserving label counts and lengths. Table 9 shows silver supervision substantially improves the 0.6B model, approaching strong closed-source LLM baselines and clearly surpassing an untuned 32B counterpart. Untuned local Qwen rows in Table 9 use direct joint two-shot prompting, while API rows use direct joint prompting. The Qwen3-0.6B-LoRA rows use fine-tuned direct generation with true silver labels or the within-video shuffled-label control. Silver labels for fine-tuning are generated locally by full DORA with Qwen3-32B. Shuffled training collapses performance, confirming that gains come from learning true discourse structure and topic alignment.

4.7 Multimodal Enhancement: Effect of Slide Integration

MOSAIC-G’s matched slide resources reflect realistic lecture consumption: learners integrate speech with visual material supplying key terms, definitions, and structural cues. To validate its utility, we test whether slide content improves knowledge-point extraction—generating topics substantively taught.

We compare two conditions: (1) **Transcript-only** uses transcript alone; (2) **Transcript+Slides** adds OCR-extracted slide text, without utterance-to-page labels. For each, Qwen3-32B generates a knowledge-point list; a head-to-head LLM judge compares the best-matching predictions with gold knowledge points. This isolates slide-derived text’s contribution while holding the model and evaluation protocol constant. Full details appear in Appendix I.

Method	Infer-Time	SRS ($\uparrow$)		Time-SRS ($\uparrow$)		WinDiff ($\downarrow$)		P_k ($\downarrow$)	
		Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard
<i>Independent classification baseline (local)</i>									
BERT-base-chinese	5min	0.29	0.28	0.30	0.29	0.95	0.96	0.60	0.55
<i>Direct joint two-shot prompting (local)</i>									
Qwen3-0.6B	7min	0.18	0.18	0.18	0.18	0.46	0.51	0.39	0.42
Qwen3-4B	12min	0.60	0.58	0.60	0.59	0.36	0.39	0.28	0.30
Qwen3-8B	18min	0.68	0.67	0.68	0.67	0.33	0.37	0.25	0.26
Qwen3-14B	22min	0.81	0.80	0.81	0.80	0.28	0.33	0.17	0.18
Qwen3-32B	51min	0.82	0.82	0.82	0.82	0.28	0.33	0.15	0.17
<i>Direct joint prompting (API)</i>									
GPT-4o-mini	–	0.67	0.67	0.68	0.67	0.33	0.37	0.20	0.21
Gemini-2.0-flash-lite	–	0.76	0.76	0.77	0.77	0.30	0.35	0.19	0.19
Gemini-2.0-flash	–	0.86	0.86	0.86	0.86	0.18	0.20	0.11	0.12
GPT-4o	–	0.92	0.91	0.92	0.91	0.17	0.21	0.08	0.10
<i>Fine-tuned with silver labels (local)</i>									
Qwen3-0.6B-LoRA	7min	0.85	0.83	0.85	0.84	0.22	0.22	0.16	0.17
<i>Ablation with shuffled labels (local)</i>									
Qwen3-0.6B-LoRA	7min	0.14	0.14	0.14	0.14	0.40	0.45	0.36	0.41

Table 9: Silver label quality validation on MOSAIC-G. Local models are run on a single NVIDIA RTX 4090 GPU; API models are accessed via commercial endpoints with inference time not directly comparable. BERT is an independent classifier. Untuned local Qwen base models use direct joint two-shot prompting; API models use direct joint prompting. Qwen3-0.6B-LoRA uses silver-label fine-tuned direct generation, and the shuffled-LoRA row is the corresponding shuffled-label direct-generation control.

Outcome	IT	PP	PA	PN	EP	MM	All
Transcript-only	24 (20.9%)	45 (17.5%)	48 (14.1%)	57 (22.3%)	29 (18.7%)	21 (12.5%)	224 (17.4%)
Transcript+Slides	52 (45.2%)	129 (50.2%)	168 (49.4%)	121 (47.3%)	65 (41.9%)	56 (33.3%)	591 (45.8%)
Ties	32 (27.8%)	72 (28.0%)	119 (35.0%)	68 (26.6%)	54 (34.8%)	86 (51.2%)	431 (33.4%)
Neither	7 (6.1%)	11 (4.3%)	5 (1.5%)	10 (3.9%)	7 (4.5%)	5 (3.0%)	45 (3.5%)

Table 10: Effect of slide-text integration on knowledge point extraction. For each gold knowledge point, we compare the best-matching predictions from Transcript-only vs. Transcript+Slides using an LLM judge; the slide condition uses OCR-extracted slide text, not utterance-to-page labels. Bold indicates best; underline indicates worst (except Neither). Transcript+Slides wins $2.6\times$ more often, with a net advantage of +367 across 1,291 gold knowledge points. Both extraction and judging use Qwen3-32B.

Results and analysis. Table 10 shows consistent gains across all 6 MOSAIC-G courses. At the aggregate level, Transcript+Slides wins 45.8% versus 17.4% for Transcript-only, with 33.4% ties and 3.5% “Neither” outcomes. This corresponds to a net advantage of +367 better-matched points across 1,291 gold knowledge points. The low “Neither” rate confirms that slides improve extraction quality rather than merely recovering from baseline failures: slides anchor terminology and disambiguate closely related knowledge points that sound similar in spoken form. These results demonstrate the value of slide text for lecture understanding beyond what the transcript alone can provide.

Error Analysis. The main error modes on MOSAIC-G are confusable topic pairs, boundary-adjacent non-topical utterances, Stage-1 propagation, and decoupled pipelines; detailed evidence and analysis are provided in Appendix J.

5 Conclusion

We introduced MOSAIC, a large-scale Chinese MOOC corpus for topic-guided segmentation with 179 courses, 14,942 videos, 7,355 hours, and strong annotation quality ($\kappa = 0.87$). MOSAIC comprises MOSAIC-G (gold benchmark) and MOSAIC-S (silver training data), uniquely combining predefined topic inventories, explicit non-topical utterances, and matched slide resources within a unified schema. Our experiments show that sibling-augmented inventories substantially degrade performance, that silver supervision from DORA enables a 0.6B model to outperform its untuned 32B counterpart, and that incorporating slide text consistently benefits knowledge-point extraction, a lecture-comprehension proxy. Future work includes fine-grained utterance-to-slide-page grounding, end-to-end multimodal segmentation, and cross-lingual extension to broaden the corpus’s applicability across educational contexts.

6 Limitations

Gold benchmark scale. The manually annotated gold benchmark covers 6 courses. While it spans diverse disciplines and supports reliable evaluation, it represents only a small fraction of the full corpus.

Assumption of a predefined inventory. MOSAIC assumes a predefined topic inventory is available in advance (e.g., an outline or knowledge-point list). This matches many realistic instructional settings, but it does not cover scenarios where no inventory exists. MOSAIC’s hierarchical metadata and course-level knowledge graphs provide a basis for future work on automatic inventory generation, which can then be refined using DORA.

Modeling bias in decoding constraints. The near-monotonic assumption is a good fit for typical lectures that progress through an outline, but it may penalize legitimate backtracking or interleaving between topics during natural instruction.

ASR and metadata noise. ASR errors and incomplete metadata can negatively affect alignment quality. Although multimodal signals can help mitigate this issue, robustness to severe noise remains a challenge in real-world deployment settings.

Slide grounding granularity. MOSAIC records video-level links to matched slide resources, but it does not include manual utterance-to-slide-page annotations. Our slide-integration experiments therefore use OCR-extracted slide text as video-level context where available and permitted. Fine-grained utterance-to-page grounding is left as an important extension for future work.

Language coverage. MOSAIC is currently Chinese-only, which may limit immediate applicability for broader multilingual research. Extending the benchmark to additional languages is a natural next step for broader cross-lingual evaluation.

7 Ethics and Broader Impact

MOSAIC is derived from publicly accessible educational content. The public release focuses on annotations, metadata, resource links, code, and documentation; it does not redistribute original videos, full slide decks, slide-page images, or slide OCR text. Access to linked source resources and use of source-derived fields remain subject to applicable permissions and platform policies; we recommend

appropriate access control and redaction practices to mitigate privacy risks.

MOSAIC is intended to support research on interpretable and reliable educational NLP. By explicitly modeling predefined topic inventories, non-topical lecture utterances, and matched slide resources, the benchmark encourages systems that better reflect how real instruction is delivered and consumed. Potential positive impacts include improved lecture navigation, retrieval, and summarization tools that can lower barriers to learning. As with any large-scale dataset, there is a risk of misuse (e.g., generating misleading educational content); we therefore emphasize research-only use and responsible deployment practices.

Acknowledgements

This work was supported by Project 62276178 under the National Natural Science Foundation of China, the Key Project 23KJA520012 under the Natural Science Foundation of Jiangsu Higher Education Institutions, Project 24BTQ002 under the National Social Science Foundation of China, Project 22YJCZH091 under the Humanities and Social Science Fund of the Ministry of Education, the Jiangsu Higher Education Reform Research Project 2025JGYB594, and the Priority Academic Program Development of Jiangsu Higher Education Institutions. They are also with Jiangsu Key Lab of Language Computing, Suzhou, 215123, P. R. China.

References

- Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. [SECTOR: A neural model for coherent topic segmentation and classification](#). *Transactions of the Association for Computational Linguistics*, 7:169–184.
- Doug Beeferman, Adam Berger, and John Lafferty. 1999. [Statistical models for text segmentation](#). *Machine Learning*, 34(1–3):177–210.
- Hang Chen, Chao-Han Huck Yang, Jia-Chen Gu, Sabato Marco Siniscalchi, and Jun Du. 2025. [MISP-meeting: A real-world dataset with multimodal cues for long-form meeting transcription and summarization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15479–15492, Vienna, Austria. Association for Computational Linguistics.
- Freddy Y. Y. Choi. 2000. [Advances in domain independent linear text segmentation](#). In *1st Meeting of*

the North American Chapter of the Association for Computational Linguistics.

- Jacob Eisenstein. 2009. [Hierarchical text segmentation from multi-scale lexical cohesion](#). In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 353–361, Boulder, Colorado. Association for Computational Linguistics.
- Jacob Eisenstein and Regina Barzilay. 2008. [Bayesian unsupervised topic segmentation](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.
- Steffen Freisinger, Philipp Seeberger, Thomas Ranzemberger, Tobias Bocklet, and Korbinian Riedhammer. 2025. [Towards multi-level transcript segmentation: LoRA fine-tuning for table-of-contents generation](#). In *Interspeech 2025*, pages 276–280.
- Marti A. Hearst. 1997. [Text tiling: Segmenting text into multi-paragraph subtopic passages](#). *Computational Linguistics*, 23(1):33–64.
- Pei-Yun Hsueh, Johanna D. Moore, and Steve Renals. 2006. [Automatic segmentation of multiparty dialogue](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273–280, Trento, Italy. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Anna Kazantseva and Stan Szpakowicz. 2014. [Hierarchical topical segmentation with affinity propagation](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 37–47, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. [Text segmentation as a supervised learning task](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.
- Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. 2019. [What Should I Learn First: Introducing LectureBank for NLP education and prerequisite chain learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681.
- Salima Mdhaffar, Yannick Estève, Antoine Laurent, Nicolas Hernandez, Richard Dufour, Delphine Charlet, Geraldine Damnati, Solen Quiniou, and Nathalie Camelin. 2020. [A multimodal educational corpus of oral courses: Annotation, analysis and case study](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4293–4301, Marseille, France. European Language Resources Association.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. [HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.
- Lev Pevzner and Marti A. Hearst. 2002. [A critique and improvement of an evaluation metric for text segmentation](#). *Computational Linguistics*, 28(1):19–36.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. [How2: A large-scale dataset for multimodal language understanding](#). In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.
- Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. [COIN: A large-scale dataset for comprehensive instructional video analysis](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1207–1216. Computer Vision Foundation / IEEE.
- Masao Utiyama and Hitoshi Isahara. 2001. [A statistical model for domain-independent text segmentation](#). In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 499–506, Toulouse, France. Association for Computational Linguistics.
- Rose E Wang, Pawan Wirawarn, Kenny Lam, Omar Khattab, and Dorottya Demszky. 2024. [Problem-oriented segmentation and retrieval: Case study on tutoring conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12654–12672, Miami, Florida, USA. Association for Computational Linguistics.
- Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2025. [Multimodal fusion and coherence modeling for video topic segmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17577–17593, Vienna, Austria. Association for Computational Linguistics.
- Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2020. [MOOCCube: A large-scale data repository for NLP applications in MOOCs](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3135–3142, Online. Association for Computational Linguistics.

Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wenzheng Feng, Wei Xu, Shulin Cao, Kaisheng Zeng, Zijun Yao, Lei Hou, Yankai Lin, Peng Li, Jie Zhou, Bin Xu, Juanzi Li, Jie Tang, and Maosong Sun. 2021. [MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4643–4652. ACM.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. [Towards automatic learning of procedures from web instructional videos](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 7590–7598. AAAI Press.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David F. Fouhey, Ivan Laptev, and Josef Sivic. 2019. [Cross-task weakly supervised learning from instructional videos](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3532–3540. Computer Vision Foundation / IEEE.

A DORA Prompt Templates and Training Details

We implement DORA as a two-stage prompting pipeline. Stage 1 (Topic List Refinement) consists of two sequential prompts: (i) conservative reconstruction that filters, reorders, and deduplicates the candidate list by retaining one representative candidate, and (ii) reflection-based correction that audits and fixes over-pruning or over-keeping. Stage 2 (Segmentation and Alignment) performs joint segmentation and topic assignment at the subtitle-line level. Across all prompts, the transcript is treated as the only ground-truth evidence source, and the model is constrained to output **JSON only** to enable deterministic parsing.

Input formatting (shared). Given a video transcript as a list of subtitle lines $\{u_i\}_{i=0}^{N-1}$, we linearize it with explicit indices:

```
[0] u_0
[1] u_1
...
```

Candidate or finalized topic inventories are formatted as indexed items (e.g., C0, C1 for candidates; K0, K1 for finalized topics).

Stage 1, Step 1: Conservative reconstruction.

Goal: From a potentially noisy candidate list, keep the topics that are substantively covered in the transcript and order them by appearance time. **Design:** recall-first (“when uncertain, keep”), transcript-as-evidence, JSON-only output.

System prompt (Step 1).

You are an expert in course syllabus analysis, skilled at identifying which topics are truly taught in a lecture transcript.

Background:

You will receive (1) a full lecture transcript with line indices and (2) a candidate topic list that may contain noise, redundancy, or wrong order.

Task:

Analyze the transcript and decide whether each candidate topic is substantively taught, then output a filtered list ordered by the time it appears in the transcript.

Keep a topic if ANY holds:

- The lecturer discusses it continuously for ≥ 5 lines, OR
- Gives definitions/formulas/principles/steps, OR
- Provides concrete examples/cases, OR
- Performs comparisons/analysis.

Remove a topic if ANY holds:

- No corresponding explanation can be found, OR
- Only mentioned briefly (< 3 lines) with no substantive content.

Critical principle:

Recall-first. If uncertain, KEEP rather than remove.

Output:

Return a JSON array ordered by appearance time:

```
[
  {"text": "<original title>", "
   source_id": <candidate index>},
  ...
]
```

- text: keep the original candidate title unchanged
- source_id: the candidate index (C0, C1, ...)

Output JSON only.

The scalar `source_id` records the representative candidate retained after deduplication; it does not encode a multi-source merge.

User prompt (Step 1).

Lecture transcript:
{TRANSCRIPT}

Candidate topics:
{CANDIDATE_TOPICS}

Filter the substantively taught topics and output them in teaching order.

Stage 1, Step 2: Reflection & finalization. Goal: Audit Step 1 for false deletions (missed topics) and false keeps (spurious topics), then output a final topic list. **Design:** bidirectional check (recover vs. delete), transcript evidence, maintain correct temporal order.

System prompt (Step 2).

You are a teaching-quality auditor responsible for verifying whether an AI-produced syllabus is accurate.

You have:

- (1) Lecture transcript (the only evidence),
- (2) Original candidate list,
- (3) Step-1 kept topics,
- (4) Step-1 removed topics.

Tasks:

- 1) Check removed topics for false deletion:
If a removed topic has substantive coverage (≥ 3 consecutive lines), restore it.
- 2) Check kept topics for false keep:
If a kept topic is not substantively covered or only name-dropped, delete it.
- 3) Ensure the final list is ordered by actual teaching time.

Output:

Return the finalized JSON array:

```
[
  {"text": "<original title>", "
    source_id": <candidate index>},
  ...
]
```

Output JSON only.

User prompt (Step 2).

Lecture transcript:
{TRANSCRIPT}

Original candidate list:
{ORIGINAL_CANDIDATES}

Step-1 kept topics:
{STEP1_KEPT}

Step-1 removed topics:
{STEP1_REMOVED}

Audit Step 1 and output the final topic list.

Stage 2: Segmentation & alignment. Goal: Segment the transcript into contiguous utterances and assign each segment either a topic id ($0 \dots M - 1$) or -1 for greetings, transitions, wrap-ups and irrelevant speech. **Constraints:** (i) near-monotonic progression over topics (rare short back-reference allowed), (ii) continuity-first (avoid fragmentation), (iii) explicit non-topical guidelines, (iv) cover all lines without overlap.

System prompt (Step 3).

You are a course transcript segmentation expert. You will receive:

- (1) A full transcript with line indices, and
- (2) A finalized, time-ordered topic list ($K_0, K_1, \dots$).

Task:

Partition the transcript into contiguous segments and label each segment with a topic id or -1 .

Core principles:

- 1) Near-monotonicity: topic ids should mostly increase with time ; occasional short review is allowed; frequent or large back-jumps are not.
- 2) Continuity-first: keep each topic as one or a few contiguous blocks; avoid scattering.
- 3) -1: greetings, roll-call, equipment checks, explicit transitions ("next we..."), closing remarks, previews, unrelated chat, or non-lecture text.
Do NOT mark as -1: fillers inside a topic, minor repetition, examples supporting the current topic, or content that explains the current topic.
- 4) Outline precedence: if the outline explicitly includes "Introduction"/"Summary" topics, map those parts to the corresponding topic id instead of -1.

Output:

Return a JSON array of segments:

```
[
  {"start_line_index": 0, "
    end_line_index": 9, "topic_id":
    -1},
  {"start_line_index": 10, "
    end_line_index": 45, "topic_id
    ": 0},
  ...
]
```

Requirements:

- Closed intervals (inclusive ends)
- Segments are contiguous, non-overlapping, and cover all lines 0..N-1
- topic_id in {-1, 0..M-1}

Output JSON only.

User prompt (Step 3).

```
Lecture transcript (N lines,
  indexed 0..N-1):
{TRANSCRIPT}
```

```
Final topic outline (M topics, K0..K
```

```
{M-1}):
{FINAL_TOPICS}
```

Segment the transcript and label each segment with the corresponding topic id or -1.

A.1 Decoding Configuration

Table 11 summarizes decoding settings for DORA stages across different model backends.

Backend	Temp.	Max Tokens	Retries
GPT-4-turbo	0.0	4096	5
GPT-4o-mini	0.0	4096	5
Qwen3-32B (local)	0.0	4096	–

Table 11: Decoding configuration. For Responses API, reasoning effort is set per stage: high for Stage 1, low for Stage 2. The GPT configurations apply to the LessonLink experiments only. The Qwen3-32B configuration applies to the DORA runs in Tables 6–7 and to MOSAIC-S silver-label generation; Table 9 uses the row-specific backbones and protocols reported there.

For models supporting the Responses API, we use `reasoning.effort` to control inference-time compute: high for topic refinement (Stage 1) where careful deliberation improves recall, and low for segmentation (Stage 2) where the task is more mechanical given a clean topic list.

A.2 LoRA Fine-Tuning Configuration

Table 12 details the hyperparameters for fine-tuning Qwen3-0.6B with LoRA on MOSAIC-S.

Output format. We use a triplet representation for training efficiency: each sample’s target is a JSON array of [topic_id, start_line, end_line] tuples, which can be deterministically converted to/from per-utterance labels. This reduces sequence length compared to outputting N individual labels while preserving full alignment information.

A.3 Annotation Cost Analysis

Tables 13 and 14 summarize the cost of generating silver annotations using DORA on the 173 additional courses beyond MOSAIC-G.

The two-stage pipeline issues approximately 3 LLM calls per video. Silver generation ran locally: ~44,124 calls in total to a locally deployed Qwen3-32B on 4×NVIDIA RTX 3090 (24 GB), BF16, at

Hyperparameter	Value
Base model	Qwen3-0.6B
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Max sequence length	8192
Batch size	1
Gradient accumulation	16
Effective batch size	16
Learning rate	2×10^{-4}
Warmup ratio	0.03
Weight decay	0.0
Epochs	3
Precision	BF16
Train/val split	90% / 10%
Eval frequency	Every 50 steps
Checkpoint selection	Best eval loss

Table 12: LoRA fine-tuning hyperparameters.

~45 seconds per video and ~184 hours of wall-clock time. For reference, replicating the same calls through a commercial API at GPT-4o-mini pricing (\$0.15/1M input, \$0.60/1M output tokens) would cost approximately \$200 — still far below manual annotation.

B Running Example: Correct vs. Plausible Wrong Segmentation

This appendix provides a concrete running example from the Personality Theory lecture (Pavlov) to illustrate the difference between correct segmentation under a predefined topic inventory and plausible but wrong topic-free segmentation. In particular, this lecture contains an explicit slide reference (“you can see it in the figure”), highlighting that matched slide resources can provide useful complementary context for grounded understanding.

B.1 Refined Topic List (Captions)

- 1 (IGNORE) Non-topical utterances (opening remarks, transitions, closing)
- 0 Pavlov’s Classical Conditioning and Its Significance
- 1 Unconditioned Reflex: Definition, Properties, and Typical Examples

Metric	Value
Total videos processed	14,708
Total LLM calls (3 prompt calls across 2 stages)	~44,124
Backbone	Qwen3-32B (locally deployed)
Hardware	4×NVIDIA RTX 3090 (24 GB), BF16
Avg. processing time/video	~45 sec
Total wall-clock	~184 hours

Table 13: Actual local silver-annotation run.

Metric	Value
Assumed pricing	GPT-4o-mini, \$0.15 / \$0.60 per 1M in/out
Estimated replication cost	~\$200.00

Table 14: Reference estimate for API-based replication (cost not incurred).

- 2 Pavlov’s Experiment and the Discovery of Classical Conditioning
- 3 Diagrammatic Formation Process of Classical Conditioning

Noisy Topics (Not Covered in This Lecture):

- Operant Conditioning
- Extinction & Generalization
- Second-Order Conditioning

B.2 Transcript (ASR Subtitles) with Segment Annotations

-1 (Opening / Recap) S0: Okay, we just mentioned Chapter 5.

S1: Behaviorism and social learning personality theory.

S2: It mainly includes Pavlov, Watson, Skinner, S3: and Bandura’s views on personality theory.

S4: Next, I will introduce them one by one.

Segment A → Topic 0 (Pavlov’s Classical Conditioning and Its Significance) S5: Section 1: Pavlov’s classical conditioning.

S6: Pavlov was a famous Russian physiologist.

S7: During his physiological research,

S8: he discovered an important psychological phenomenon,

S9: later called classical conditioning.

S10: Regarding classical conditioning,

S11: its core manifestation is what we call the conditioned reflex.

S12: What does the conditioned reflex mean for animals and humans?

S13: We know this is a very important learning phenomenon.

⋮

S18: In classical conditioning,

S19: what exactly is the conditioned reflex that is produced?

-1 (Transition) S20: Before we understand conditioned reflexes, you need to

S21: briefly review the prerequisite for how a conditioned reflex forms.

Segment B → Topic 1 (Unconditioned Reflex: Definition, Properties, and Typical Examples)

S22: That prerequisite is the unconditioned reflex.

S23: Conditioned reflexes are generally built upon unconditioned reflexes.

S24: So what is an unconditioned reflex?

S25: We can understand it as an organism's inborn reflex,

S26: also called a species reflex.

⋮

S40: The most common one is the sucking reflex.

⋮

S52: The second example is the grasp reflex.

S53: It is an unconditioned reflex shown after birth,

S54: using the palm to grasp objects.

S55: **You can see the details in the figure.**

⋮

S83: Some unconditioned reflexes are highly complex.

S84: Conditioned reflexes can form on the basis of these complex unconditioned reflexes.

S85: As you can imagine, their level may be higher,

S86: and their manifestations may be more complex.

Segment C → Topic 2 (Pavlov's Experiment and the Discovery of Classical Conditioning)

S87: The earliest research on these more complex conditioned reflexes

S88: came from the Russian physiologist Pavlov, whom we just mentioned.

S89: More than a hundred years ago, the Russian physiologist Pavlov

S90: mainly conducted physiological research.

S91: One of his research projects was to

S92: observe dogs' salivation while eating.

⋮

S96: When a plate of food is placed in front of a dog, the dog eats.

S97: During eating,

S98: the dog naturally salivates.

S99: Pavlov called this phenomenon an unconditioned reflex.

⋮

S107: Pavlov found another phenomenon:

S108: whenever the plate of meat appeared before the dog,

S109: the food was accompanied by a bell sound.

S110: After repeatedly presenting these two stimuli together,

S111: the bell alone, even without food,

S112: for example before the plate of meat was brought out,

S113: would still make the dog salivate.

⋮

S118: Pavlov and his assistants first observed this phenomenon

S119: when the feeder's footsteps sounded outside the laboratory door.

S120: The feeder carrying the food had not yet entered,

S121: and the dog had not seen the food;

S122: after merely hearing the familiar footsteps,

S123: the dog began to salivate in advance.

⋮

S128: The dog began to salivate,

S129: a phenomenon that is quite interesting.

Segment D → Topic 3 (Diagrammatic Formation Process of Classical Conditioning)

S130: Pavlov's research can be illustrated with a vivid diagram.

S131: Let us borrow this diagram from prior materials and take a look.

S132: The figure vividly depicts how a dog

S133: moves from an unconditioned reflex to a conditioned reflex.

S134: We call the whole process conditioning,

S135: or what later became known as classical conditioning.

S136: Start from the upper-left corner.
 S137: In the first panel, a plate of food appears before a dog,
 S138: and the dog salivates while eating.
 ⋮
 S160: The third step is the key step:
 S161: a sound is presented whenever the dog is fed, or just before feeding.
 S162: In other words, whenever food appears, it is accompanied by a sound;
 S163: alternatively, the sound comes first,
 S164: and then the food appears.
 ⋮
 S185: This is what the fourth panel shows:
 S186: after food and sound have repeatedly appeared together,
 S187: the dog no longer needs to eat or see the food;
 S188: it merely hears the bell,
 S189: and begins to salivate.
 ⋮
 S206: Once this temporary neural connection forms,
 S207: the neutral stimulus becomes a meaningful conditioned stimulus,
 S208: and a conditioned response is elicited.
 S209: This is how classical conditioning is established.

-1 (Closing) S210: Thus, in Pavlov’s research,
 S211: he presented an important idea
 S212: about classical conditioning.

B.3 Correct Alignment (Gold)

Table 15 shows the correct (gold) alignment for this lecture.

ID	Topic Name	Range
-1	(Opening)	S0–S4
0	Classical Conditioning: Significance	S5–S19
-1	(Transition)	S20–S21
1	Unconditioned Reflex (def. & examples)	S22–S86
2	Pavlov Experiment & Discovery	S87–S129
3	Formation Process (diagram-based)	S130–S209
-1	(Closing)	S210–S212

Table 15: Correct alignment (gold) for the Pavlov example. The slide-referred cue at S55 motivates multimodal grounding: text-only systems may treat it as filler, while matched slide resources provide useful context for interpreting such utterances.

B.4 Why This Example Demonstrates the Value of Predefined Topics

Using the Pavlov lecture excerpt in Appendix B, this example demonstrates several practical challenges that a predefined topic inventory is well-suited to address:

- 1. Clear non-topical utterances handling:** The opening recap (S0–S4), the bridging transition (S20–S21), and the closing wrap-up (S210–S212) contain recap/organizational remarks rather than topic-specific instruction. These spans should be labeled as non-topical (-1). Without explicit non-topical modeling, as in prior benchmarks, they are easily absorbed into adjacent topical segments, contaminating both boundary detection and topic assignment.
- 2. Semantic distinctiveness among covered topics:** The four topical segments are conceptually separable and align cleanly with the predefined inventory: Segment A/Topic 0 introduces classical conditioning and its significance (S5–S19); Segment B/Topic 1 defines and exemplifies unconditioned reflexes as prerequisites (S22–S86); Segment C/Topic 2 presents Pavlov’s dog experiment and the discovery phenomenon (S87–S129); Segment D/Topic 3 explains the diagrammatic formation process step by step (S130–S209). This separation makes cross-topic mis-assignments interpretable and correctable under an inventory-guided alignment objective.
- 3. Hard negative resistance to inventory noise:** The initial recap explicitly mentions multiple behaviorist/social-learning figures (Pavlov, Watson, Skinner, Bandura; S2–S3), while the topical portion of the excerpt develops only Pavlov-related content (S5–S209). In a realistic inventory, closely related but *not covered here* topics (e.g., Skinner’s operant conditioning, reinforcement schedules, Bandura’s observational learning) act as hard negatives. A topic-free baseline may drift toward these semantically adjacent distractors, whereas an inventory-aware model should learn to reject them when the transcript provides no supporting evidence beyond the brief recap.
- 4. Transition absorption:** The transition (S20–S21) previews the prerequisite needed to understand conditioned reflexes (i.e., unconditioned

reflexes) but does not itself provide definitional content. It should therefore remain -1 rather than being forced into Segment A/Topic 0 (S5–S19) or Segment B/Topic 1 (S22–S86), which would blur the Segment A/Segment B boundary and distort the subsequent alignment.

C Traditional Baseline Comparison

To contextualize our results, we compare against traditional segmentation methods and pipeline approaches. Table 16 presents segmentation-only metrics, and Table 17 shows joint segmentation-alignment results for pipeline methods that first segment then retrieve.

Traditional segmentation methods fail to capture topic-guided boundaries effectively. More critically, pipeline approaches that separate segmentation from alignment cannot recover from segmentation errors, resulting in very low joint accuracy even with strong neural retrievers like ColBERT. This validates our design choice of joint modeling in DORA.

Method	$P_k \downarrow$	WinDiff $\downarrow$
Stage-Avg	0.3995	0.6877
Stage-Max	0.4049	0.6902
TopicSeg	0.4386	0.6463
TextTiling	0.4461	0.6703
Top-10	0.6009	1.0000
Top-20	0.6013	1.0000

Table 16: Traditional Segmentation Methods Evaluation Results

Segmentation	BM25	TF-IDF	Jaccard	ColBERT
TopicSeg	0.1470	0.1470	0.0778	0.0476
Stage-Avg	0.0444	0.0431	0.0307	0.0251

Table 17: Alignment Methods Evaluation (SRS) Based on Best-Performing Segmentation Approaches

D Non-topical Utterance Ratio on MOSAIC-G

Table 18 reports the proportion of non-topical utterances and their corresponding time ratio across the 6 MOSAIC-G courses.

Course	Videos	Utterance Ratio	Time Ratio
IT	22	14.1%	16.0%
PP	52	3.3%	3.1%
PA	54	3.2%	3.2%
PN	48	3.5%	3.7%
EP	28	3.7%	3.2%
MM	30	3.3%	3.7%
All	234	4.6%	5.2%

Table 18: Ratio of non-topical utterances and time on MOSAIC-G.

E Noisy Inventory Generation

To construct noisy topic inventories for stress-testing (Section 4), we use GPT-4o-mini to generate confusing but plausible distractors. For each video, the model receives a transcript preview and the original topic list, then produces the same number of not-covered knowledge points following these guidelines:

- **Related but different scope:** e.g., “nitrogen fertilizer application methods” for a lecture on “nitrogen fertilizer effects”
- **Adjacent concepts:** e.g., “respiration” for a lecture on “photosynthesis”
- **Different granularity:** more specific sub-topics or more general parent topics
- **Prerequisites/follow-ups:** topics that typically precede or follow the actual content

The prompt template is shown below:

System: You are an expert in educational content analysis. Generate confusing but plausible knowledge points that are NOT covered in a lecture, but could easily be mistaken as being covered.

User:

```

=== TRANSCRIPT (first 3000 chars)
===
{transcript}

=== ACTUAL KNOWLEDGE POINTS COVERED
===
{captions}

```

```

=== TASK ===
Generate exactly {n} confusing
knowledge points that:
1. Are NOT covered in this
transcript
2. Are semantically related and
could be easily mistaken as
covered
3. Would be plausible topics in the
same course

Output: JSON array of strings only.

```

Generated distractors are appended to the original inventory while preserving the original label-to-topic mapping. During inference, the combined list is shuffled, and predictions are mapped back to original IDs for evaluation.

F Gold-Course Selection Protocol

We selected the six MOSAIC-G courses under the following operational criteria:

1. **Supporting-material completeness:** an instructor-provided knowledge-point outline for every video, matched slide deck(s), and a course-level knowledge graph.
2. **Outline granularity at knowledge-point level:** 4–7 topics per video (realized 4.9–6.3; Table 3).
3. **Moderate video length:** average 10–20 minutes (realized 12.4 min; Table 4).
4. **ASR quality:** manual spot-checks of sampled utterances across multiple videos, retaining only courses without systematic recognition failures.
5. **Disciplinary coverage:** exactly one course from each of six top-level categories.

MOSAIC-G is deliberately a clean benchmark: reliable gold labels are a precondition for meaningful evaluation, and the resulting agreement ($\kappa=0.87$, Table 5) validates this choice. Robustness beyond the clean condition is evaluated separately: the noisy-inventory setting (§4) stress-tests confusable candidates, and MOSAIC-S covers the remaining 173 courses without such filtering, with models fine-tuned on it transferring strongly to MOSAIC-G (Table 9). MOSAIC-G is also not uniformly easy: IT contains 14.1% non-topical utterances (Appendix D) and the most confusable terminology (largest drop under noise, Table 6).

Course	Abnormality rate (%)
EP	0.35
IT	0.34
MM	0.44
PA	0.47
PN	0.04
PP	0.01
Corpus average (14,942 videos)	0.63

Table 19: Subtitle-abnormality rate (empty, garbled, duplicated, or timestamp-inverted caption lines) per gold course vs. the corpus-wide average over all 14,942 videos.

Statistic	Value
Raw inventory size $ C $	10.90 ± 2.51
Refined inventory size $ T' $	5.80 ± 1.75
Removed at Step 1 (%)	46.67
Restored by Step-2 reflection (%)	0.00
Pairwise order-inversion rate (%)	48.70
Segments per realized topic	1.010

Table 20: Stage-1 refinement statistics, measured on the DORA runs over the 234 gold-course videos; the MOSAIC-S pipeline is identical. Segments per realized topic uses the contiguous-label-run definition: 1,302 topical runs divided by 1,289 realized (video, topic_id) pairs gives 1.010. Existing logs contain only a scalar `source_id` for the retained representative and do not encode multi-source merges, so no merge count is reported.

G Stage-1 Refinement Statistics

H Metric Definitions

This appendix gives the formal definitions of the evaluation metrics used in §4.

Segmentation metrics. We adopt two standard topic segmentation metrics: P_k (Beeferman et al., 1999) and WinDiff (Pevzner and Hearst, 2002). Both operate on a sliding window of size $k = \lfloor N/2R \rfloor$, where N is the number of utterances and R is the number of reference segments. Throughout, superscript $*$ denotes the reference segmentation, and quantities without $*$ are computed from the prediction. P_k measures the probability that two utterances k apart are incorrectly judged to be in the same segment or in different segments:

$$P_k = \frac{1}{N-k} \sum_{i=1}^{N-k} \mathbb{1}(\delta^*(i, i+k) \neq \delta(i, i+k)), \quad (1)$$

where $\delta(i, j) \in \{0, 1\}$ indicates whether utterances i and j belong to the same predicted segment, and $\delta^*(i, j)$ is defined analogously for the reference. Intuitively, P_k penalizes both false boundaries (splitting a true segment) and missed boundaries (merging distinct segments). WinDiff counts windows where the predicted and reference numbers of boundaries disagree:

$$\text{WinDiff} = \frac{1}{N - k} \sum_{i=1}^{N-k} \mathbb{1}(b^*(i, i+k) \neq b(i, i+k)), \quad (2)$$

where $b(i, j)$ is the number of boundaries between utterances i and j in the prediction, and $b^*(i, j)$ is the corresponding count in the reference. Lower values indicate better segmentation.

Alignment metrics. Segmentation metrics are insufficient because a system may place plausible boundaries but still assign segments to wrong topics, especially under confusable candidates. Following Wang et al. (2024), we adapt the Segmentation and Retrieval Score (SRS) to topic alignment. In our setting, its line-based form is instantiated as per-utterance topic-label accuracy:

$$\text{SRS} = \frac{1}{N} \sum_{j=1}^N \mathbb{1}(w_j = w_j^*), \quad (3)$$

where w_j and w_j^* are predicted and gold labels for utterance j . SRS measures utterance-to-topic alignment quality: any boundary shift or wrong topic assignment directly reduces the score.

We also report Time-SRS (Wang et al., 2024), which weights each utterance by its duration:

$$\text{Time-SRS} = \frac{1}{\sum_j \alpha_j} \sum_{j=1}^N \alpha_j \cdot \mathbb{1}(w_j = w_j^*), \quad (4)$$

where $\alpha_j = t_j^{\text{end}} - t_j^{\text{start}}$ is the duration of utterance j . Higher values indicate better alignment.

I Slide-Integration Evaluation Protocol

Task formulation. Knowledge point extraction serves as a proxy for lecture comprehension: a system that truly understands a lecture should be able to distill its content into a list of covered knowledge points. We compare two input conditions: (1) **Transcript-only** uses transcript alone; (2) **Transcript+Slides** additionally includes OCR-extracted slide text, without using utterance-to-page

labels. For each condition, we prompt an LLM (Qwen3-32B) to freely generate a list of knowledge points that are substantively taught (with definitions, explanations, or examples), without providing any candidate list or length constraint. The generated lists are then compared against the gold knowledge points (instructor-provided outlines) to measure how well each condition covers the taught content.

Evaluation protocol. We adopt a head-to-head comparative protocol: for each gold knowledge point, we find the best-matching prediction from each condition, then ask an LLM judge (Qwen3-32B, the same backbone as the extraction model) to decide which match is more accurate (winner), whether they are equivalent (tie), or whether neither provides a satisfactory match. This paired comparison focuses the judgment on relative correctness, avoiding calibration issues that arise when scoring predictions independently.

J Detailed Error Analysis

We consolidate the main error modes observed on MOSAIC-G. (1) *Confusable topic pairs*: most alignment errors involve semantically adjacent candidates within a discipline; the effect concentrates in IT, whose dense economics terminology yields the largest degradation under the sibling-augmented inventory (−9 percentage points, Table 6). (2) *Boundary-adjacent non-topical utterances*: disagreements cluster at segment edges where openings and transitions blend into topical content. The Soft–Hard SRS gap is small for five courses, but IT is a clear exception (about 5 pp under Clean and 4 pp under Noisy; Table 6). (3) *Stage-1 propagation*: without refinement, spurious or missing inventory items propagate into Stage 2 as incorrect assignments; Stage 1 reconstruction without reflection recovers +2.7 pp average SRS (Table 7), with the benefit tracking the degree of inventory–lecture mismatch (largest on IT and PA, mixed on EP). (4) *Decoupled pipelines*: segment-then-retrieve baselines collapse ($\text{SRS} < 0.15$, Appendix C) because early segmentation errors are unrecoverable at the alignment stage, which motivates joint segmentation–alignment decoding.