Cut to the Chase: Training-free Multimodal Summarization via Chain-of-Events

Supplementary Material

A. Generalization Test

To systematically assess the generalization capability of Chain-of-Events (**CoE**), we conduct extensive cross-domain experiments across other seven diverse MMS benchmarks: BLiSS [18] (instructional videos), MM-AVS [14] and XMSMO [44] (news reports), SoccerNet [34] (sports broadcasts), Summ [35] (TV series), and TIB [16] and VISTA [30] (lecture videos). These datasets exhibit substantial domain variations in content structure, visual semantics, and temporal dynamics.

For each dataset, we train **MLASK** [23] and **MMSum** [39] baselines using their official implementations, then evaluate them on all remaining datasets without fine-tuning or domain adaptation. In contrast, **CoE** operates in a purely zero-shot manner across all benchmarks, requiring no task-specific training. As shown in Figures A6–A3, the results exhibit highly consistent trends:

- **In-domain overfitting of supervised methods.** Models trained on individual datasets perform well within their training domain but experience substantial performance drops when evaluated on different domains. For instance, a model trained on BLiSS shows strong results on instructional videos but struggles with sports footage from SoccerNet or scripted dialogues from SummScreen. This degradation occurs regardless of which dataset is used for training, revealing a fundamental limitation: existing fusion-based methods tend to overfit to domain-specific characteristics such as visual style, narrative pacing, or vocabulary distribution.
- **Consistent zero-shot transfer with CoE.** Unlike supervised baselines, our training-free framework maintains steady performance across all datasets, while baseline methods exhibit unstable patterns when tested out-of-domain. This stability arises naturally from **CoE**'s design, which adapts to different video genres through hierarchical event modeling rather than relying on dataset-specific patterns learned during training.
- **Competitive performance even without in-domain supervision.** **CoE** also achieves *comparable or even better* scores than supervised baselines on several datasets (e.g., MM-AVS, BLiSS, Summ). These observations indicate that **CoE's** event-centric architecture, which combines hierarchical event graph construction, cross-modal grounding, event-evolution reasoning, and lightweight style adaptation, provides a strong inductive bias that alleviates the need for task-specific supervision while still yielding competitive performance on target domains.

In summary, these cross-domain experiments show that existing supervised MMS models are strongly tied to their training domains, whereas **CoE** delivers stable zero-shot performance across heterogeneous benchmarks. By relying on an event-centric architecture instead of dataset-specific supervision, **CoE** offers a more domain-agnostic solution for MMS and remains competitive even on datasets where other methods are explicitly trained.

B. Prompts

B.1. Hierarchical Event Graph (HEG) Construction

CoE first decomposes the input text to construct a hierarchical event graph, which contains three layers: a global event layer, a sub-event layer, and an entity-relation layer. The global event captures the overall theme of the narrative, while sub-events decompose the global event into semantically coherent components. We use the following prompt to jointly infer the main event and its sub-events:

> Analyze this video transcript or article and provide:
>
> 1. A main event summary in one sentence that captures the high-level essence of the entire video. Keep it concise, preferably under 50 words.
>
> 2. Determine if the video can be meaningfully divided into sub-events (maximum 3). If yes, provide concise one-sentence summaries for each sub-event. Keep the granularity appropriate - don't make the sub-events too specific or detailed.
>
> Return your response in JSON format:
> If the transcript CANNOT be divided into sub-events:
> {"main_event": "your main summary here", "sub_events": [main event]}
> If the transcript CAN be divided into sub-events:
> {"main_event": "your main summary here", "sub_events": ["sub-event 1", "sub-event 2", ...]}
>
> Here are two examples of good main event summaries:
> - {Example 1}
> - {Example 2}
>
> The following is the video transcript: {input text}

Then, for each sub-event, **CoE** refines its representation at the entity-relation layer by extracting an event-specific set of typed entities and organizing them into a subgraph. Concretely, we prompt the MLLM to identify all entities that are directly relevant to the given sub-event, including

(i) *persons* involved in or mentioned within the event, (ii) *locations* where the event takes place, (iii) *organizations* such as companies, institutions, or teams, and (iv) salient *objects/items* that play a role in the event. All entities are required to be explicitly grounded in the transcript and deduplicated. These event-specific entities serve as the nodes of the entity-relation subgraph, which is later used to provide fine-grained semantic anchors for cross-modal grounding and reasoning.

Please extract relevant entities related to the specified event from the provided video transcript. Entities should include people, locations, organizations, and objects/items. Return the results in JSON format with the following structure:
{
"person": ["name 1", "name 2", "name 3"],
"location": ["place 1", "place 2", "place 3"],
"organization": ["org 1", "org 2", "org 3"],
"item": ["item 1", "item 2", "item 3"]
}

Instructions:
1. Only include entities directly related to the specified event
2. Remove duplicates and normalize entity names
3. For people, use full names when available
4. For organizations, use official names rather than abbreviations when possible
5. Ensure all extracted entities are actually mentioned in the transcript
6. If no entities of a certain type are found, return an empty array for that category

Event: {event}
Video transcript: {text}

In summary, the HEG construction module transforms unstructured textual input into a structured semantic scaffold by sequentially decomposing the narrative into global events, sub-events, and fine-grained entity sets. By organizing information across these three hierarchical layers, we capture both the high-level thematic evolution and the precise entity-level details necessary for reasoning. This constructed graph serves as a robust domain-agnostic prior, providing the essential semantic anchors that guide the subsequent cross-modal grounding and event evolution analysis.

B.2. Cross-modal Spatial Grounding (CSG)

In the CSG module, we first perform a sub-event alignment procedure grounded in the HEG. Given an input video, we uniformly sample a fixed number of frames and partition them into shot-level clips, each of which encapsulates a coherent local temporal context. Guided by the HEG, every clip is then associated with its most relevant sub-event. The prompt employed in this process is defined as follows:

Please analyze the {domain} scene in this set of images and determine which of the following sub-events this video clip belongs to.

Return the result in JSON format with three fields:
- "idx": the chosen sub-event number (starting from 0),
- "sub-event": the description of the chosen sub-event,
- "reasoning": a short explanation of why this sub-event was selected based on the video content.

The JSON format must be: { "idx": number, "sub-event": "event description", "reasoning": "short explanation"}

Sub-events: {sub-event list}

Following the assignment of clips to specific sub-events, we proceed to fine-grained spatial grounding. In this step, the goal is to construct a visually grounded subgraph by verifying which entities are visible in the current clip and identifying their interactions. The prompt designed to extract these visual entity-relation triples is provided below:

Extract entity relationships from the video by following these steps:

1. List visible entities: What people, organizations, or locations appear?
2. Identify actions/relationships: What connections exist between entities?
3. Match with subgraph: Find corresponding triples in the provided subgraph
4. Verify relevance: Ensure matches are reasonable (direct or contextual)

If no exact match exists, select the most contextually relevant triple.

Subgraph: {subgraph}

Output format: {"reasoning": "your analysis", "triples": [{"from": "David Warner", "relation": "celebrated his century at", "to": "Coogee Oval"}]}

Through this two-stage prompting strategy, the CSG module effectively transforms the raw video stream into a sequence of visually grounded subgraphs. By enforcing both temporal alignment with sub-events and spatial verification of entity relations, we ensure that the reasoning process is anchored in concrete visual evidence rather than relying solely on textual priors. This yields fine-grained, reliable correspondences between visual observations and textual concepts, providing a robust structured representation for the subsequent EER module.

B.3. Event Evolution Reasoning (EER)

Building on the Video Clip Aggregation step 3.4, the EER module focuses on capturing the dynamic evolution of the narrative. Instead of treating segments in isolation, this module analyzes the causal and temporal dependencies between them.

Specifically, for each aggregated temporal segment, the model is provided with the global textual context, the segment's visual content, and a structural comparison between the current entity-relation graph and that of the preceding segment. By contrasting the added and removed entities or relations across adjacent segments, the model can distinguish between mere visual redundancy and genuine narrative transitions . This setup allows the model to explicitly track the trajectory of the event and to identify how entities persist, interact, or change over time, instead of merely describing static scenes, which in turn provides a more stable event backbone for the subsequent summary generation module.

The prompt used to reason about these transitions and generate the trajectory description is as follows:

Based on the reference transcript, the given event and sub-event graph, and the newly identified entities and relations, extract the most relevant information from the transcript. Then analyze the event trajectory and generate a concise description in no more than 100 words.

STRICT OUTPUT FORMAT (no extra text):

Event trajectory: {analysis of event progression}
Description summary: {concise summary}

Inputs:
- Reference transcript: article
- Sub-event: sub-event
- Subgraph: subgraph
- New entities and relations: {New entities and relations}

B.4. Domain-adaptive Summary Generation (DSG)

The final module, DSG, operates in two distinct phases to synthesize the final output. First, we construct a comprehensive initial summary by aggregating the sequence of event trajectory descriptions produced by the EER module. Rather than simply concatenating these descriptions, we prompt the MLLM to consolidate the key narrative developments, causal transitions, and entity interactions into a unified and coherent text.

This initial draft serves as a content-heavy backbone, ensuring that all salient multimodal information extracted during the reasoning process is preserved. The prompt used to generate this event-centric initial summary is as follows:

Generate a {domain} summary following these steps:
1. Identify key information:
   - Main event and participants from the article
   - Timeline and locations from scene descriptions
   - Critical facts and outcomes
2. Structure the summary:
   - Lead with the most important fact
   - Follow with supporting details
   - Keep similar length to examples
3. Match example style:
   - Concise, fact-focused sentences
   - No commentary or interpretation
   - Include names, numbers, specific details

Examples:
{Example 1}
{Example 2}
{Example 3}

Overall event: {total event}
Sub-events: {sub-events}
Entities and relations: {entities and relations}
Scene descriptions: {scene descriptions}
News Article: {article}

While the initial summary captures the factual content, it may lack the specific linguistic nuances of the target domain. To address this, the second phase employs a lightweight style adaptation mechanism. We retrieve a small set of reference summaries from the target domain to serve as stylistic exemplars. The MLLM is then prompted to rewrite the initial summary, aligning its tone, phrasing, and discourse structure with these exemplars while strictly maintaining content fidelity. The prompt for this style refinement process is provided below:

Rewrite the text by thinking through these steps:
1. Analyze examples' characteristics:
- Sentence structure (short, fact-dense)
- Information density (3-5 key facts per example)
- Length range (estimate word count from examples)

2. Extract key facts from input text:
- Who, what, when, where, why
- Remove redundancy and commentary

3. Reconstruct in example style
- Use compact sentences
- Maintain factual accuracy
- Match length to example

Examples (style and length reference):
{Example 1}
{Example 2}
{Example 3}

Text: {initial summary}

Dataset	Domain	# Train	# Val	# Test
VIEWS	News	141.0 K	1.6 K	1.6 K
MM-AVS	News	1.8 K	–	0.4 K
XMSMO	News	4.4 K	0.3 K	0.3 K
TIB	Video Lectures	7.3 K	0.9 K	0.9 K
VISTA	Academic	14.9 K	1.8 K	1.8 K
BLiSS	Social Media	11.0 K	2.5 K	2.5 K
SoccerNet	Sports	0.5 K	0.06 K	0.1 K
Summ	Entertainment	5.2 K	3.0 K	3.0 K

Table A1. Summary of datasets used in our experiments, grouped by domain, along with the number of samples in the training, validation, and test splits.

By decoupling content synthesis from stylistic refinement, the DSG module ensures robust generalization across diverse benchmarks. This separation allows **CoE** to maintain high factual accuracy while flexibly adapting its linguistic output to match the distinct conventions of news, sports, or instructional videos without requiring domain-specific fine-tuning.

C. Dataset

To ensure a comprehensive evaluation across distinct video genres, we conduct experiments on eight diverse MMS benchmarks. These datasets cover five primary domains: news broadcasting, instructional videos, sports commentary, academic presentations, and entertainment narratives.

This extensive selection allows us to validate the model's performance on varied content structures and linguistic styles. Specific details for each benchmark are described as follows:

- **VIEWS.** Representing the news broadcasting domain, VIEWS [3] is a large-scale benchmark originally designed for entity-aware video captioning. Derived from the M^2E^2R [4] corpus, it mitigates the loose alignment issues typical of news datasets by utilizing LLMs to generate ground-truth captions from shot-specific event descriptions. Unlike generic benchmarks, VIEWS requires models to explicitly identify named entities and interpret dynamic news contexts from visual cues. To adapt this dataset for the MMS task, we leverage the official video descriptions from YouTube to serve as the textual input.
- **MM-AVS.** A full-scale benchmark in the news domain, MM-AVS [14] is curated from CNN and Daily Mail archives. It distinguishes itself through comprehensive modality coverage, providing a rich set of inputs for each news story, including articles, videos, audio, transcripts, images, and captions. This dataset challenges models to synthesize heterogeneous information from these diverse streams into concise summaries, effectively addressing the lack of holistic audio-visual integration in prior benchmarks.
- **XMSMO-News.** Targeting the "Too Long; Didn't Watch" (TL;DW) scenario, XMSMO-News (XMSMO) [44] establishes the task of *extreme* MMS using content derived from BBC News. Unlike standard benchmarks, it challenges models to condense lengthy video-document pairs into an ultra-compact output comprising a single cover frame and a one-sentence summary. This rigorous constraint serves to evaluate the model's capacity to distill the most salient visual and textual semantics into a minimal representation, effectively addressing the challenge of information overload.
- **TIB.** A benchmark for the academic domain, TIB [16] focuses on abstractive summarization of long-form video-conference recordings. Containing over 9,100 lectures with an average length of 37 minutes, it demands robust long-context modeling capabilities. Unlike standard datasets, TIB requires processing technically dense narratives and aligning asynchronous modalities—including speech, visual gestures, and presentation slides—to produce accurate summaries.
- **VISTA.** Centered on the domain of scientific communication, VISTA [30] is a large-scale benchmark tailored for video-to-text summarization of academic discourse. It comprises 18,599 aligned pairs of conference recordings and their canonical paper abstracts, curated from premier AI venues such as the ACL Anthology, ICML, and NeurIPS. By targeting the gap in processing technical terminology and scientific visual elements, VISTA challenges models to synthesize long-form multimodal content (averaging 6.8 minutes) into highly structured, factually dense summaries. To adapt this dataset for the MMS task, we leverage Whisper [41] to generate transcripts from the videos.
- **BLiSS.** Sourced from the Behance platform, BLiSS [18] serves as a premier benchmark for the livestreaming and instructional domains. It comprises over 13,000 video-transcript pairs specifically focused on creative artistic processes. Distinguished by its raw, untrimmed nature, BLiSS features content with extensive durations and slow-paced temporal dynamics, presenting a distinct challenge compared to traditional short-video datasets. Furthermore, it provides temporally aligned transcripts and dual-modality ground truth (key-frames and key-sentences), enabling rigorous evaluation of cross-modal alignment in long-horizon, realistic streaming environments.
- **SoccerNet-Caption.** SoccerNet-Caption (SoccerNet) [34] is a large-scale benchmark specifically designed for dense video captioning within full-length soccer

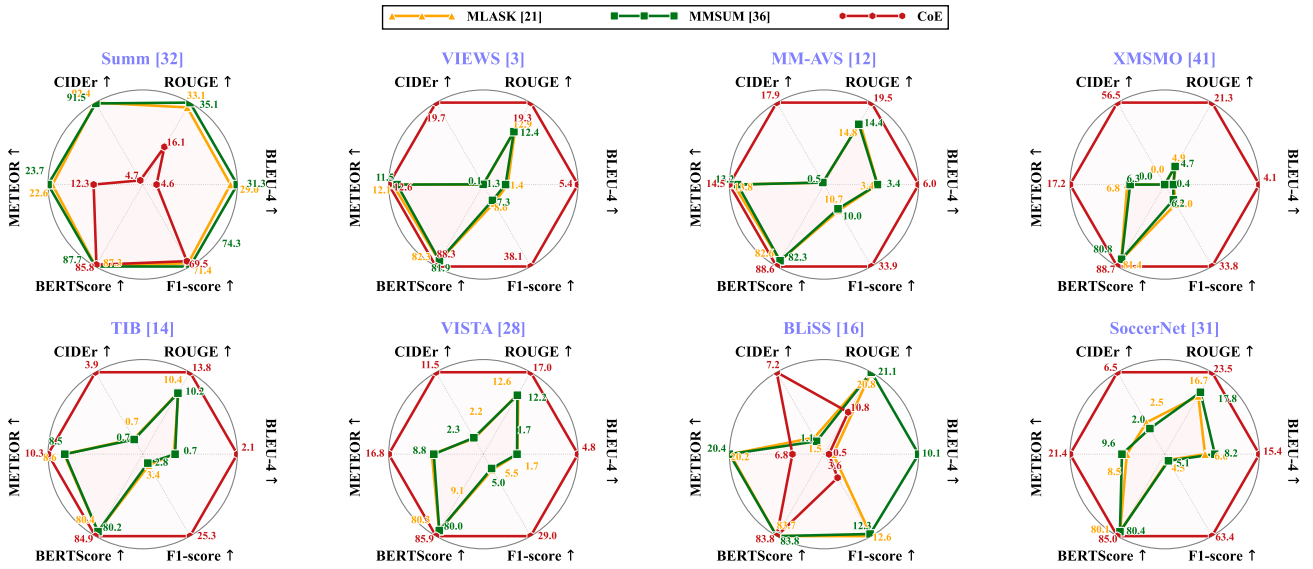


Figure A1. **Motivating Experiments on Summ.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on Summ [35], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

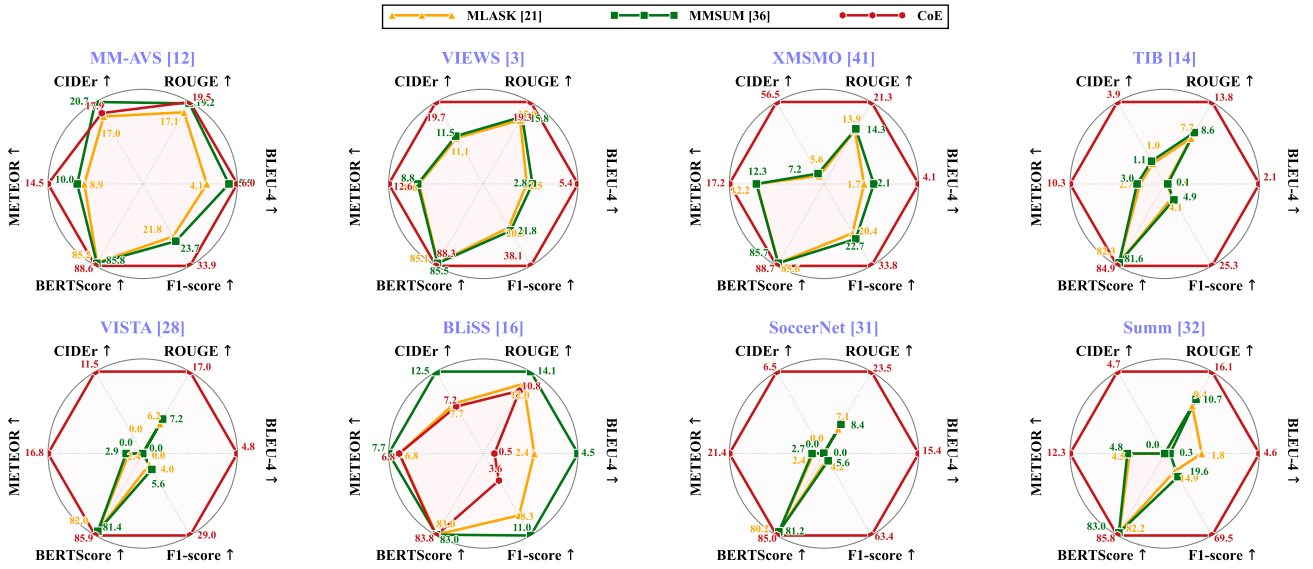


Figure A2. **Motivating Experiments on MM-AVS.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on MM-AVS [14], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

broadcasts. It consists of 36,894 timestamped commentaries distributed across 715.9 hours of untrimmed video footage, sourced from 471 games in major European leagues. Distinct from traditional benchmarks that rely on temporal intervals, this dataset introduces the Single-anchored Dense Video Captioning (SDVC) task, challenging models to generate rich, emotionally charged, and factually precise narratives anchored to specific game events. With average video durations exceeding 45 minutes, it rigorously tests a model's ability to handle extensive temporal contexts and domain-specific terminology in a live broadcast setting.

- **SummScreen^{3D}.** Curated for the entertainment narrative domain, SummScreen^{3D} (Summ) [35] constitutes a multimodal extension of the dialogue-centric SummScreen corpus [7]. It features 4,575 full-length TV episodes (av-

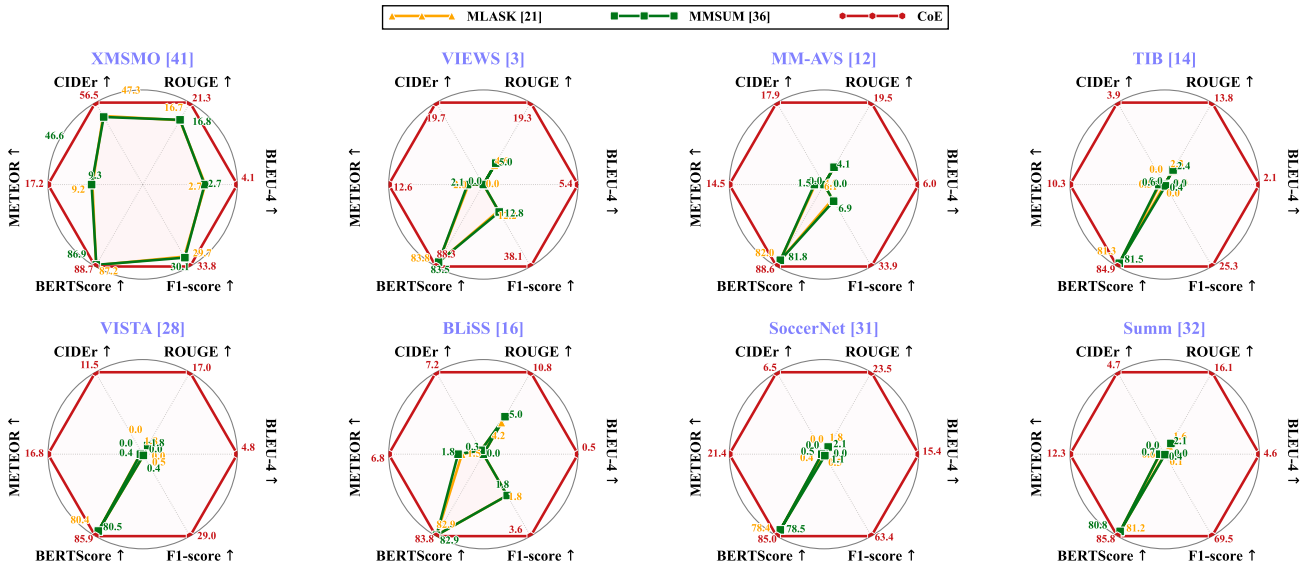


Figure A3. **Motivating Experiments on XMSMO.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on XMSMO [44], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

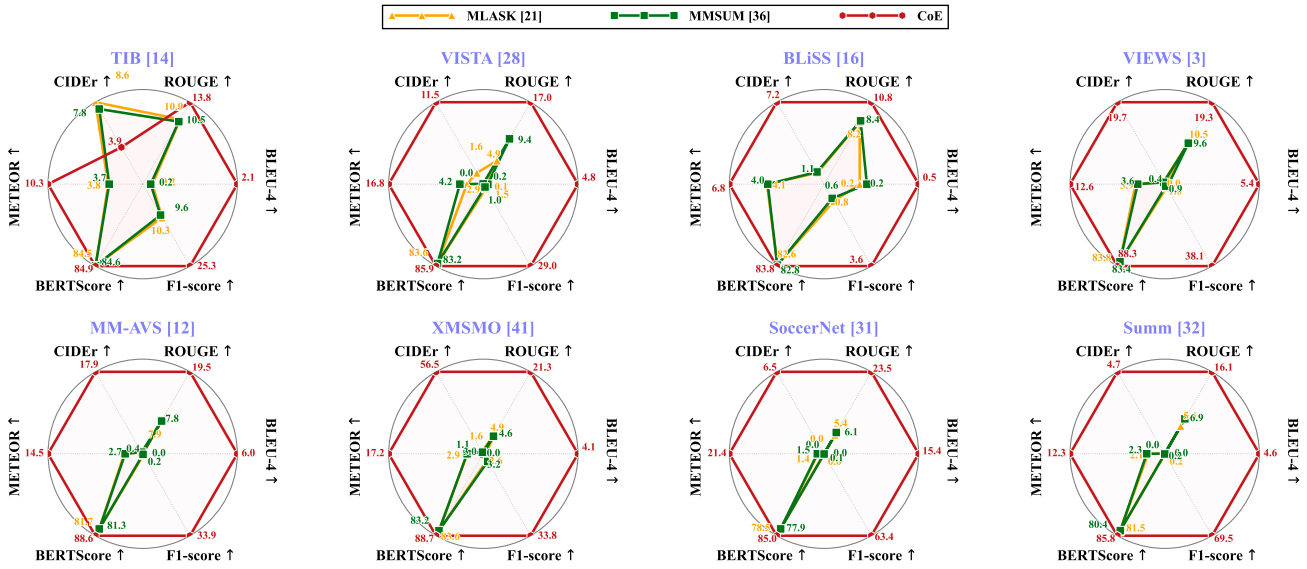


Figure A4. **Motivating Experiments on TIB.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on TIB [16], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

eraging 40 minutes) paired with transcripts and abstractive summaries derived from soap operas. This benchmark establishes a rigorous testbed for long-form video-to-text summarization, compelling models to synthesize visual and acoustic cues, such as character emotions and scene dynamics, with extensive dialogue to resolve long-range plot dependencies.

Detailed statistics regarding the training, validation, and testing splits for all datasets are summarized in Table A1.

D. Implementation Details

D.1. Implementation Details of Baselines

To thoroughly evaluate the efficacy and superiority of our proposed **CoE** framework, we employ two distinct classes of baseline models in our comparative experiments. The

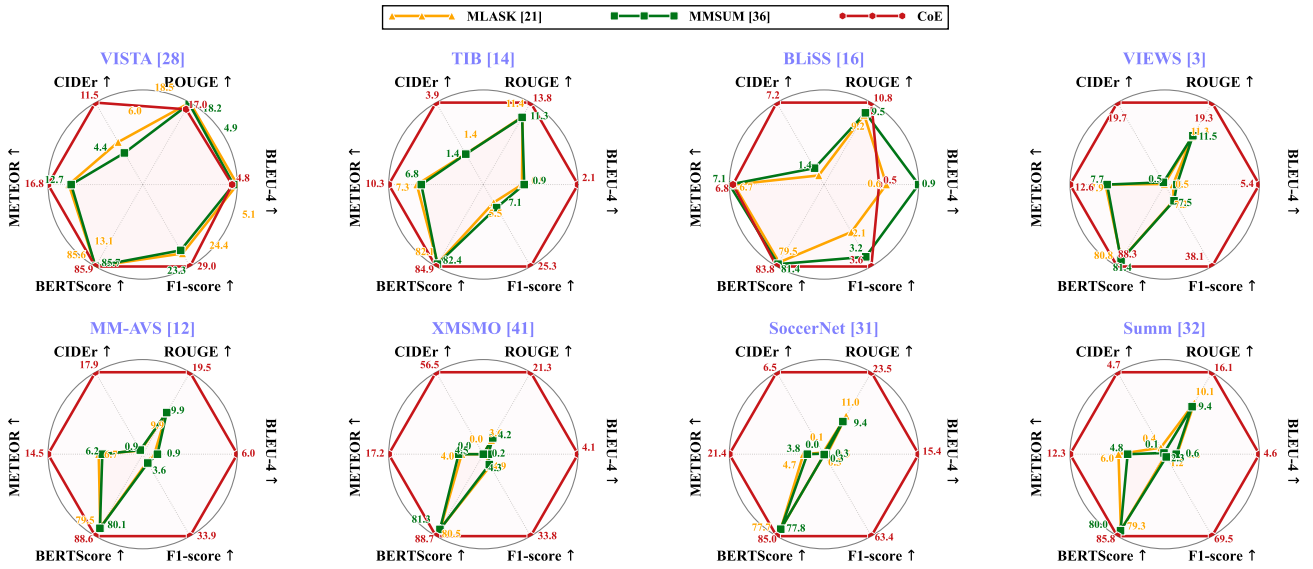


Figure A5. **Motivating Experiments on VISTA.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on VISTA [30], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

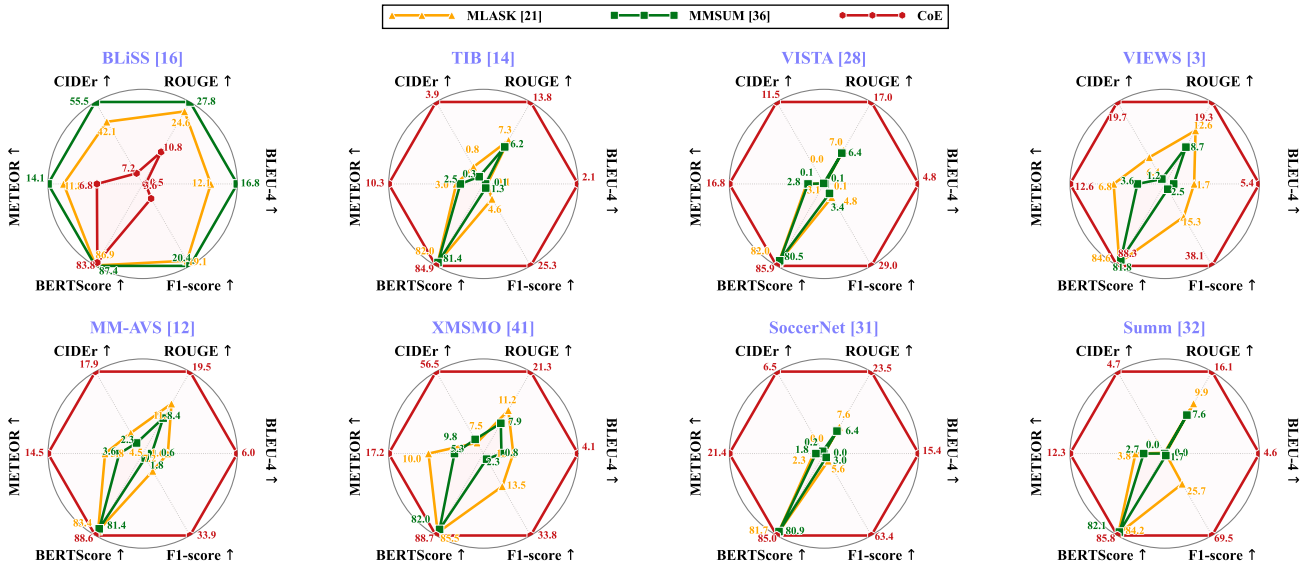


Figure A6. **Motivating Experiments on BLiSS.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on BLiSS [18], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

first class comprises state-of-the-art traditional MMS methods, including MLASK [23] and MMSum [39]. These models are included primarily to examine the performance of our training-free **CoE** framework under scenarios of domain transfer and to rigorously test its generalization capability, offering a crucial performance benchmark against domain-dependent counterparts. The second class of baselines focuses on the video Chain-of-Thought (CoT) reasoning mechanisms, specifically encompassing TCoT [2], CoF [15], ViTCoT [59], and CoS [20]. This selection provides a direct comparison point, allowing us to quantify the superiority of the **CoE** framework in structured event reasoning and event flow modeling over existing CoT strategies. The implementation details are as follows:

- **MLASK.** We fine-tune MLASK [23] using its official implementation. In addition to generating textual summaries, the model is designed to select a representative video frame as a cover image. Since our datasets

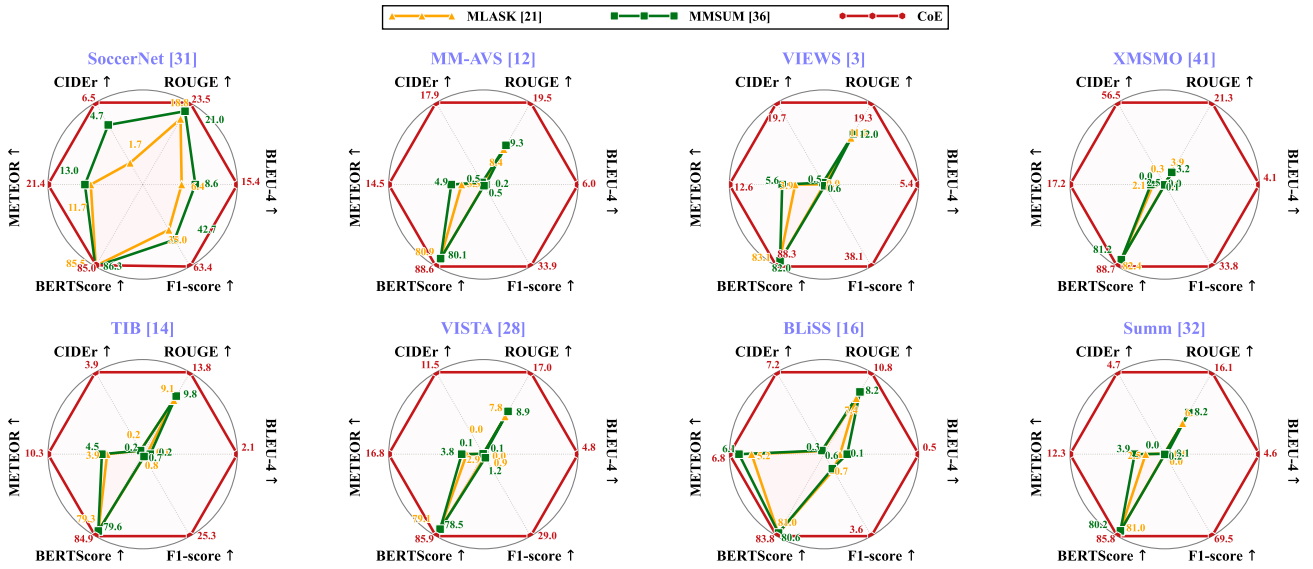


Figure A7. **Motivating Experiments on SoccerNet.** Existing MMS models (e.g., **MLASK** [23] and **MMSum** [39]) achieve strong in-domain results when trained on SoccerNet [34], but their performance drops sharply under domain shift. In contrast, our **training-free CoE** framework generalizes effectively across diverse datasets, maintaining stable zero-shot performance without task-specific training or adaptation.

do not contain ground-truth cover annotations, we construct the cover images as follows. For VIEWS, MM-AVS, XMSMO, SoccerNet, Summ, and BLiSS, we use CLIP [40] to identify the frame with the highest semantic similarity to the transcript sentences and set it as the cover image. For TIB and VISTA, we directly use the first frame as the cover, as it typically corresponds to the title slide or agenda.

- **MMSum.** We deploy MMSum [39] following its official implementation, strictly adhering to its standard protocols for feature extraction. Similar to MLASK, MMSum jointly generates a textual summary and a visual thumbnail. To compensate for the lack of native thumbnail annotations in our datasets, we mirror the data processing strategy used for MLASK: we utilize the same CLIP-based alignment to derive pseudo-ground-truth thumbnails for general video domains, while defaulting to the initial frame as the target for presentation-oriented content (TIB and VISTA).
- **TCoT.** Due to the absence of an official open-source implementation, we re-implement Temporal Chain of Thought (**TCoT**) [2] strictly following the algorithmic design described in the original paper. To ensure a fair comparison, we instantiate the framework using the same MLLM backbone as our method (Qwen2.5-VL-7B-Instruct [5]). Regarding hyperparameter settings, we adhere to the protocols reported in the original work: we adopt the *Dynamic-Segment* inference strategy with a segment size of $s = 64$ and a segment count of $l = 12$. Input videos are sampled at 1 fps, and the maximum context window is constrained to 32K tokens to match the original experimental budget.
- **CoF.** We use the official `CoF-8B` checkpoint from Chain-of-Frames (**CoF**) [15], which is fine-tuned on the InternVL3-8B [60] backbone. Regarding video processing, we uniformly sample frames at 1 fps and set the maximum input length to 64 frames. During inference, we adapt the original CoT prompt to the summarization task. Specifically, we modify the instruction to elicit frame-aware reasoning traces followed by a comprehensive video summary, rather than a question-specific answer.
- **ViTCoT.** We re-implement Video-Text Interleaved Chain-of-Thought (**ViTCoT**) [59] on the same Qwen2.5-VL-7B-Instruct backbone as our method, strictly following the two-stage Video-Text Interleaved CoT paradigm and the prompt templates described in the original paper. In the first stage, the model performs standard text-only CoT reasoning conditioned on the video and transcript. In the second stage, we sample the raw video at 1 fps and use CLIP [40] to select a small set of high-similarity frames, which are then interleaved into the intermediate reasoning trace to simulate the oracle key-video used in ViTCoT. The maximum number of key frames per video and other decoding hyperparameters follow the configuration reported by Zhang et al. [59].
- **CoS.** We adapt Chain-of-Shot prompting (**CoS**) [20] to our MMS setting as a training-free baseline. Following the original design, we use an LLaVA-based [58] classifier to perform binary video summarization over mo-

saiced frame groups, and LongVA as the downstream summarization backbone. For each video, we uniformly sample frames at 1 fps with at most 32 frames, group every four consecutive frames into a composite image for binary relevance prediction, and then construct positive and negative shot sequences from the resulting binary codes. The original QA-oriented prompts in CoS are rewritten into summarization-style instructions so that LongVA directly generates textual summaries conditioned on the selected shots, while other hyperparameters follow the settings in the original CoS paper.

D.2. Implementation Details of CoE

Our **CoE** framework uses the Qwen2.5-VL-7B-Instruct as the backbone and 1 fps with a maximum of 72 frames. The frame segment size is set to 6 frames and max merged segments up to 5 (30 frames). Our proposed **CoE** framework is implemented based on the powerful Qwen2.5-VL-7B-Instruct, leveraging its strong multimodal understanding and long-context capabilities. As a training-free inference framework, the primary implementation configuration lies in efficient video sampling and structured prompt engineering. Specifically, the input video is first uniformly sampled at a rate of 1 fps, and the maximum total number of frames is capped at 72. To facilitate the hierarchical event modeling of CoE, the sampled frames are grouped into frame segments, with each segment containing 6 frames. Subsequently, we constrain the prompt context to include a maximum of 5 merged segments (equating to a total of 30 frames), which are strategically interleaved with the Hierarchical Event Graph (HEG) and the textual transcript to guide the **CoE** structured reasoning process. For decoding hyperparameters, the temperature is set to 0.1 and the maximum generated token count is fixed at 500.

D.3. Implementation Details of Ablation Studies

This subsection describes the implementation of ablations for the four modules in **CoE**. Since the pipeline is progressive, with each stage consuming the structured outputs produced by its predecessor, we adopt a bypass strategy that disables one module at a time while keeping the remaining stages functional and directly comparable under identical inference settings.

CoE – HEG (w/o Hierarchical Event Graph Construction). HEG serves as a global-to-local semantic scaffold, organizing the narrative into a hierarchy from global events to sub-events and further to entity-relation subgraphs, which in turn informs subsequent grounding and reasoning. To ablate HEG while keeping the downstream interfaces intact, we use the raw input text T as the sole textual context. Specifically, we omit the construction of the three-level HEG and its associated sub-event graphs, and instead pass T directly to later stages, where event and entity cues are extracted from T whenever needed.

CoE – CSG (w/o Cross-modal Spatial Grounding). CSG associates each video clip with a corresponding sub-event anchor and grounds entity-relation triples by verifying their visual evidence, yielding visually supported subgraphs for subsequent reasoning. In **CoE – CSG**, we remove CSG while keeping the rest of the pipeline unchanged. Concretely, we assign the text-derived subgraph produced in HEG to each clip or segment directly, without performing clip-level entity identification or relation grounding. Consequently, EER and DSG operate on ungrounded, text-only subgraphs, thereby eliminating the benefit of fine-grained visual verification.

CoE – EER (w/o Event Evolution Reasoning). EER groups semantically coherent clips into longer temporal segments and characterizes event trajectories by analyzing subgraph changes across adjacent segments. In **CoE – EER**, we disable EER and perform reasoning directly on the subgraphs matched in the previous stage, without modeling temporal evolution. Specifically, we do not compare subgraphs between neighboring segments or track the emergence and persistence of entities and relations over time. Instead, trajectory descriptions are generated solely from the subgraph(s) associated with the current clip or segment, removing explicit temporal transition modeling.

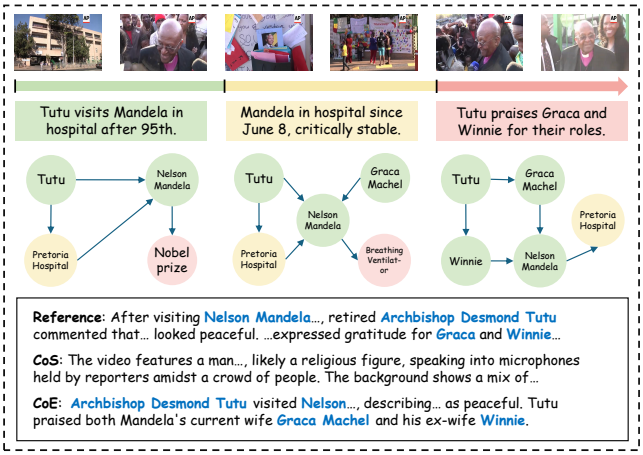
CoE – DSG (w/o Domain-adaptive Summary Generation). DSG first synthesizes an event-centric summary from the inferred trajectories and then applies lightweight style adaptation using a small set of in-domain exemplars. In **CoE – DSG**, we retain the event-centric summary synthesis but omit the style adaptation step. The initial summary is directly taken as the final output without further rewriting, isolating the contribution of domain adaptation and assessing whether stylistic refinement is necessary for strong benchmark performance.

E. Additional Experiments

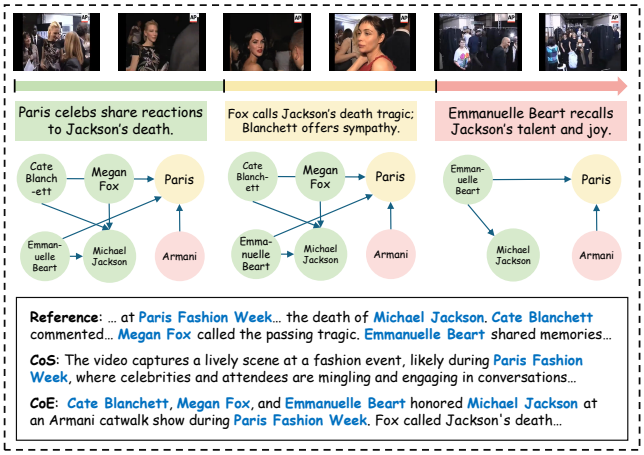
E.1. Case Study

In the main paper, we provided a detailed qualitative analysis on the Summ dataset to illustrate the effectiveness of our hierarchical reasoning framework (Section 4.6). To further examine the versatility and robustness of **CoE** across heterogeneous domains, we include additional case studies on five benchmarks: VIEWS (Figure A8), MM-AVS (Figure A9), XMSMO (Figure A10), TIB (Figure A11), and VISTA (Figure A12). These examples cover a broad range of video genres, including news broadcasts, instructional content, and scientific presentations. Similar patterns can be observed from BLiSS [18] and SoccerNet [34]. To be concise, we omit additional results here.

As shown in Figures A8 through A12, **CoE** consistently recovers the correct hierarchical event structures, aligns

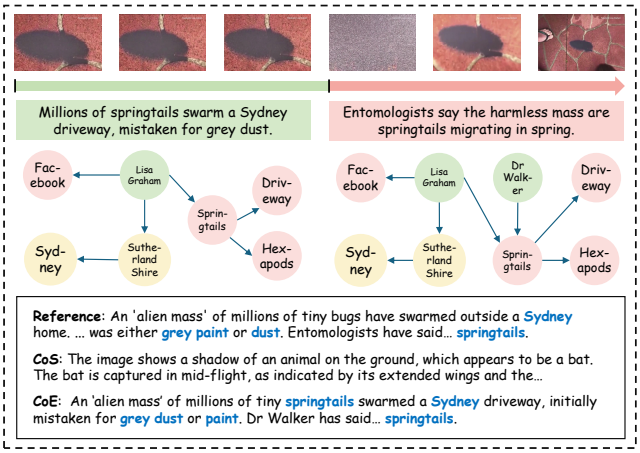


(a) **CoE** grounds news events mentions across video and news articles.

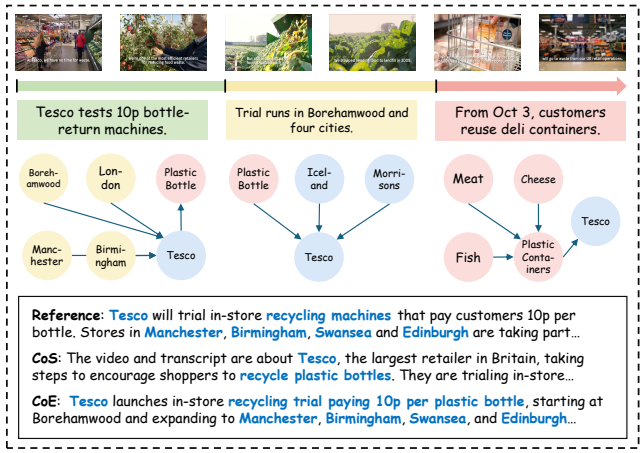


(b) **CoE** tracks event progression and produces concise news summaries.

Figure A8. **Case study on VIEWS**. (a) Guided by the HEG, **CoE** correctly links speeches by public figures to the corresponding entities and locations, aligning visual evidence with article-referenced mentions. (b) **CoE** follows the evolution of the news story across scenes and generates compact news-style summaries, while the baseline mainly enumerates local visual details without capturing the full narrative.



(a) **CoE** aligns news clips with the correct sub-events and entities.



(b) **CoE** summarizes articles by following event development.

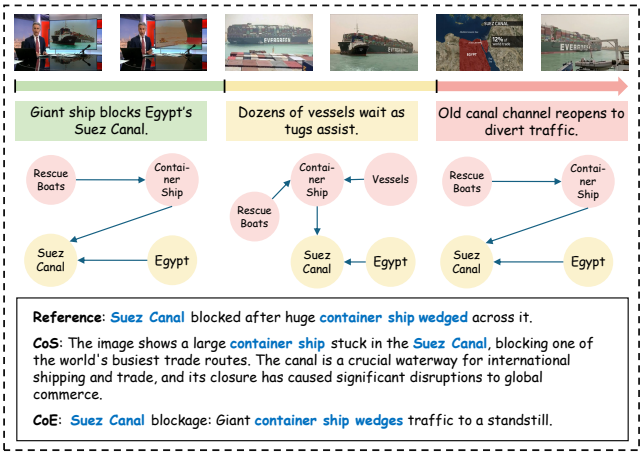
Figure A9. **Case study on MM-AVS**. (a) **CoE** uses the event graph as a scaffold to attach each news clip to the correct sub-event and to ground entities such as locations, organizations, and key objects in the scene. (b) By aggregating grounded clips along the event trajectories, **CoE** produces summaries that mirror how the news article develops, whereas the baseline remains close to scene-level description and often ignores the structure of the story.

sub-events with the corresponding visual evidence, and grounds entities and relations with fine-grained precision. Compared with baseline methods that tend to focus on local scene description, **CoE** preserves global temporal consistency and accurately captures long-range narrative development, demonstrating the utility of hierarchical event modeling for diverse video styles and discourse forms. On TIB and VISTA, we also observe that **CoS** [20] yields qualitatively reasonable summaries that are consistent with its quantitative trends on long video understanding and that it can accurately retrieve key visual details from extended sequences.

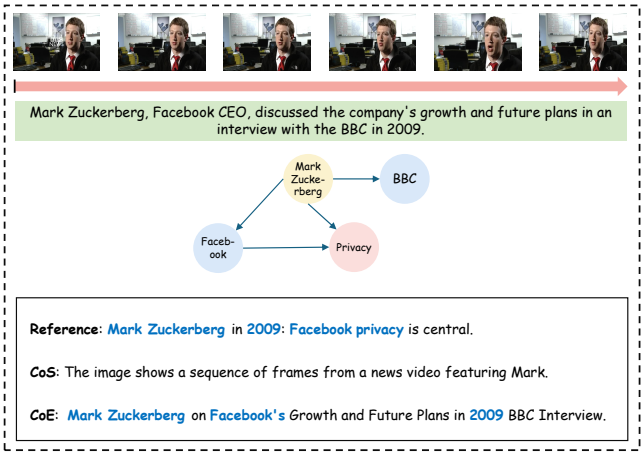
Overall, the case studies confirm that the event-centric design of **CoE** generalizes effectively across domains without task-specific fine-tuning. The generated summaries remain grounded, coherent, and stylistically appropriate, further validating the advantages of explicit structured reasoning in MMS.

E.2. Effect of Video Clip Size

To investigate the impact of temporal granularity on the reasoning capabilities of **CoE**, we conduct an ablation study on the video clip size. Specifically, we fix the total number of sampled frames per video at 72 and vary the number of frames per clip, denoted as K, within the set $\{2, 4, 6, 8\}$. This setup alters the total number of temporal segments pro-

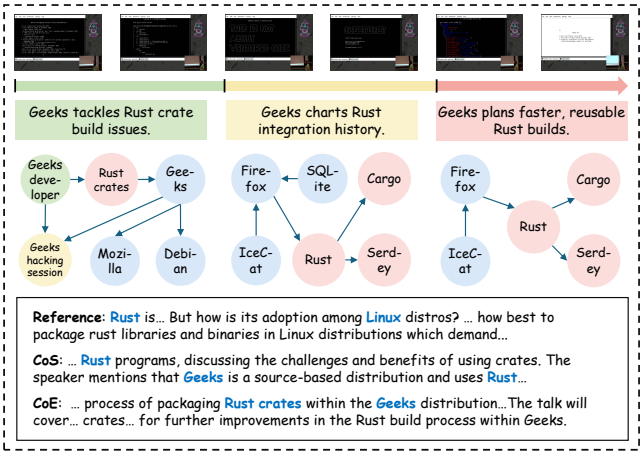


(a) **CoE** produces concise headline style summaries for extreme news cases.

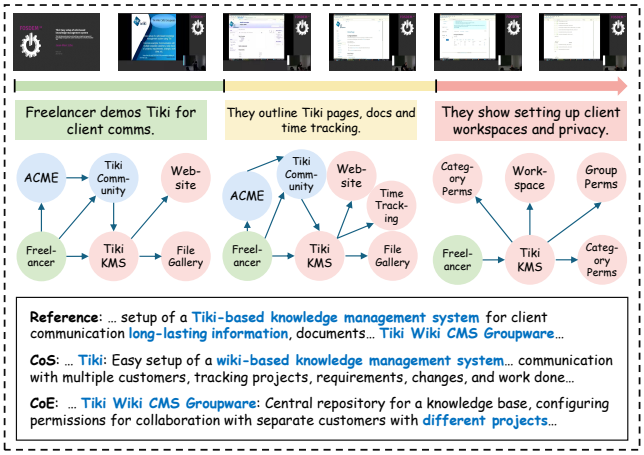


(b) **CoE** correctly identifies the main subject and interview context.

Figure A10. **Case study on XMSMO**. (a) For the Suez Canal blockage example, **CoE** compresses the multi-shot sequence into a short headline-style summary that closely follows the reference, while the baseline, which cannot exploit style exemplars, outputs a long verbose caption that reads like a literal description of the scene. (b) For the "Mark Zuckerberg" interview, **CoE** uses the hierarchical event graph and transcript to recover the correct identity and role of the speaker and to summarize the topic of the interview, whereas the baseline never names "Zuckerberg" and only provides a generic description of a news video.



(a) **CoE** grounds technical concepts and entities in long lecture recordings.



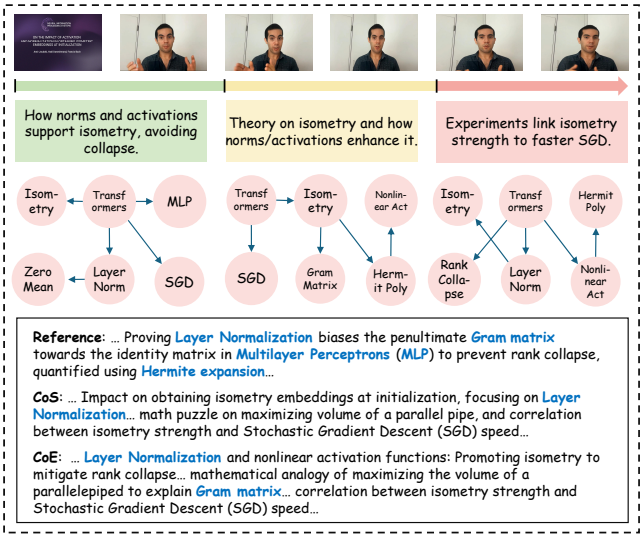
(b) **CoE** follows the structure of the talk and preserves key pedagogical steps.

Figure A11. **Case study on TIB**. (a) In technical lectures about topics such as Rust packaging or knowledge management with Tiki, **CoE** aligns clips with the corresponding sub-events and grounds abstract entities like tools, libraries, and platforms using the slide and speaker context. (b) **CoE** tracks how the speaker introduces the problem, presents solutions, and discusses future work, which leads to summaries that match the logical flow of the talk, whereas the baseline mainly reports isolated visual scenes without reconstructing the full argument.

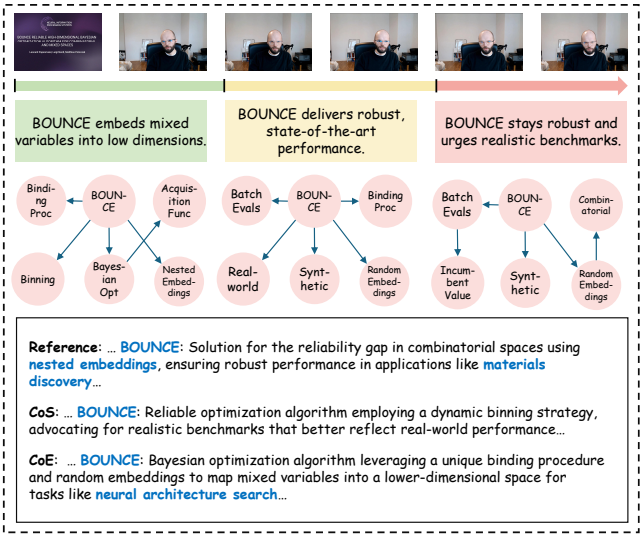
cessed by the Event Evolution Reasoning (EER) module, directly influencing the model's ability to capture local motion dynamics versus global event transitions. By holding the total information budget constant, we isolate the effect of information distribution across temporal segments.

The results are illustrated in Figure A13. We observe that a clip size of $K = 6$ consistently achieves superior or competitive performance across the majority of benchmarks. Specifically, $K = 6$ demonstrates a clear advantage on datasets requiring fine-grained motion understanding and long-term dependency modeling, such as XMSMO (21.28 ROUGE) and Soccernet (23.50 ROUGE). Furthermore, it yields the highest performance on BLiSS (10.83), significantly outperforming the $K = 2$ baseline (8.85). While $K = 8$ remains competitive in some scenarios, it does not consistently surpass the gains achieved by the 6-frame setting, and smaller clip sizes generally lag behind in performance metrics.

We attribute these findings to a trade-off between context sufficiency and information density. Smaller clip sizes (e.g., $K = 2$) tend to fragment the visual context, disrupting the continuity of actions and making it difficult for the



(a) **CoE** links slide content and spoken narration for technical talks.



(b) **CoE** follows the progression from motivation to method and findings.

Figure A12. **Case study on VISTA**. (a) For a talk on isometry and layer normalization, **CoE** uses the hierarchical event graph to segment the presentation into coherent conceptual blocks and to ground mathematical notions such as Gram matrix and stochastic gradient descent in the corresponding slide regions. (b) In the BOUNCE example, **CoE** tracks how the speaker introduces the problem, presents the algorithm based on random embeddings for mixed variables, and summarizes empirical observations, which leads to a concise summary that highlights the main scientific message. **CoS** also produces reasonable descriptions on VISTA, but its outputs are typically more generic and rely less on the detailed terminology that appears on the slides.

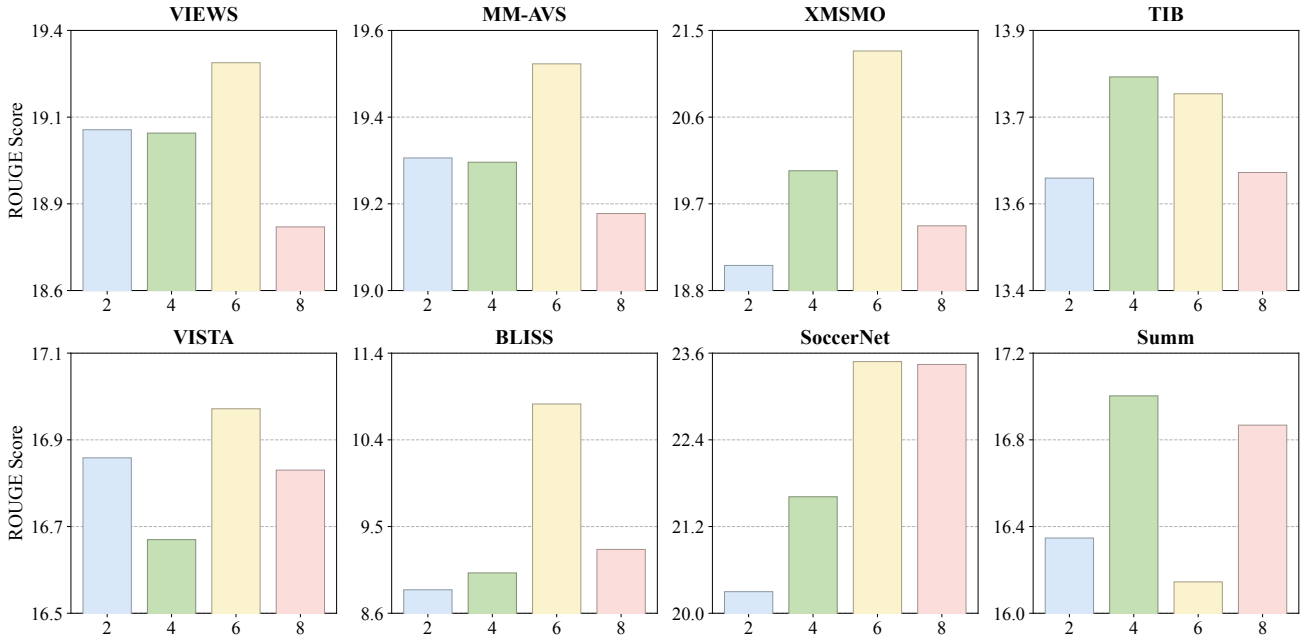


Figure A13. **Effect of Temporal Granularity.** We report the ROUGE scores across eight datasets with varying video clip sizes ($K \in \{2, 4, 6, 8\}$). While performance varies slightly across domains, a clip size of $K = 6$ yields the most robust and consistent improvements.

model to form coherent event trajectories, as evidenced by suboptimal scores on TIB and VISTA. Conversely, excessively large clip sizes (e.g., $K = 8$) may introduce visual redundancy or irrelevant background noise within a single processing unit, potentially diluting the salient features necessary for precise summarization. Therefore, we identify $K = 6$ as the optimal balance point, providing sufficient temporal context for accurate entity-relation ground-

Method	TCoT	CoF	ViTCoT	CoS	CoE
Time (s)	36.90	29.01	17.05	39.06	28.51

Table A2. **Mean runtime comparison between CoE and video CoT baselines.**

ing without overwhelming the model with redundant visual information, and thus select it as the default setting for our **CoE** framework.

E.3. Inference Time

To assess inference efficiency, we randomly sample 50 videos from each of the eight datasets and measure the end-to-end runtime per video under the same hardware and software settings for all methods. We report the average inference time across all sampled videos. As shown in Table A2, **CoE** runs in **28.51**s per video on average, making it the second fastest method overall while maintaining strong overall performance in the main comparisons.