EVA-Cap: Optimizing Audiovisual Video Captioning via Event-Centric Alignment

Jiapeng Shi[1,2*], **Qiuxia Hou**[3*], **Jiaqi Leng**[1,2], **Junke Wang**[1,2], **Yanhao Zhang**[3], **Haonan Lu**[3], **Ziyi Ye**[1,2†], **Zuxuan Wu**[1,2†]
[1]Institute of Trustworthy Embodied AI, Fudan University
[2]Shanghai Key Laboratory of Multimodal Embodied AI
[3]OPPO AI Center, OPPO Inc.

Abstract

Recent advances in Omni Language Models (OLMs) provide a unified solution for audiovisual video captioning. Nevertheless, existing approaches predominantly treat captioning as an unstructured text generation process, neglecting the inherent semantic structure of videos and leading to suboptimal fine-grained audiovisual alignment. To bridge this gap, we propose EVA-Cap, a novel framework that decomposes captions into atomic audiovisual events and transforms unstructured text supervision into event-centric alignment. We train EVA-Cap via Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) on a self-curated dataset, EVA-Data, which is built with event-centric quality control. During the GRPO phase, the model is explicitly optimized against three event-centric alignment targets: global event completeness, intra-event attribution, and inter-event synchronization. Extensive experiments demonstrate that EVA-Cap significantly outperforms existing methods in event-centric evaluation (*e.g.*, +9.0% average gain on ChronusAV), while achieving competitive performance on general audiovisual captioning benchmarks (*e.g.*, +3.3% on WorldSense). Our code is available at https://github.com/OPPO-Mente-Lab/EVA-Cap.

1 Introduction

Video captioning [2, 13, 45, 61, 54] aims to extract rich semantic content from videos and generate coherent natural language descriptions. As a fundamental task in video understanding, it underpins various downstream applications, including video retrieval [44, 38, 53, 63, 37], video reasoning [43, 42, 62, 10, 26, 31, 3], and video generation [46, 41, 39, 40, 17, 18]. Recently, advances in Transformer-based end-to-end models have revolutionized the conventional modularized video caption pipelines. Especially, Omni Language Models (OLMs) [32, 22, 59, 52, 58, 7] have driven a paradigm shift from visual-only modeling to an omni-modal paradigm that integrates audio and visual signals, enabling more effective cross-modal reasoning and stronger generalization in real-world audiovisual scenarios.

Despite its promise, achieving factually accurate and comprehensive audiovisual captioning remains difficult due to the intrinsic complexity and dense interdependence between visual and auditory content. Existing methods treat video captioning as a text generation problem for OLMs based on caption-level supervision [4, 49]. This overlooks the structured semantics inherent in videos, meaning a generated caption may be linguistically correct but misaligned with the logical structure and event-level correspondences. While some recent research has incorporated additional supervision over specific aspects (*e.g.*, dialogue [5, 33] or object attributes [21, 6]), such narrowly scoped supervision still struggles to achieve fine-grained audiovisual alignment. Consequently, a critical bottleneck stems

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

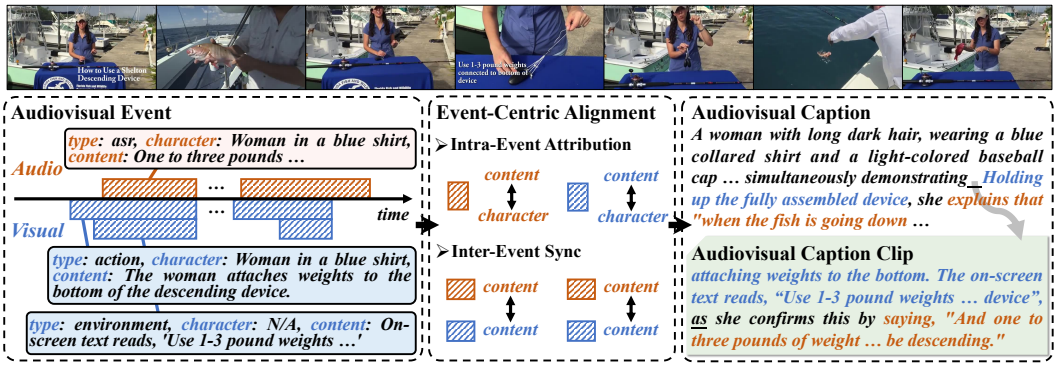


Figure 1: Illustration of the event-centric representation of audiovisual captions, where aligned audiovisual events are composed into the predicted caption.

from the lack of an effective caption representation that unifies video-grounded structural semantics with general language modeling.

To address this issue, we propose framing video captioning as assembling discrete, atomic audiovisual event sequences through joint optimization of event content and their correspondences. As illustrated in Fig. 1, a video consists of multiple audio and visual events, where each event is represented as a tuple of type, character, and content. These events can be aligned temporally to form a coherent caption clip, linked via shared semantic content and explicit synchronous connectors (*e.g.*, "as", "at the same time"). Crucially, such an event-centric perspective bridges audiovisual events and captions by modeling both intra-event attribution (i.e., whether the attributions of a single event are correct?) and inter-event synchronization (i.e., what are the temporal order and synchronization of different events?). While such an event-centric perspective provides a structured semantic understanding of videos beyond mere language modeling, effectively integrating such insights into OLMs post-training remains an open question. This gap is largely due to (1) the scarcity of high-quality audiovisual caption data with event-level fidelity, and (2) the absence of dedicated event-centric frameworks for leveraging supervision with event-level heuristics.

To this end, we propose EVA-Cap, which builds upon Qwen2.5-Omni [51] and unifies event-centric alignment within the language modeling post-training phase. Specifically, we first construct EVA-Data, a 63K audiovisual video–caption dataset, via an automatic initialization–refinement pipeline followed by event-centric quality control. In this process, timestamped audiovisual events are extracted from raw videos and subjected to per-event verification to ensure high fidelity. Furthermore, we introduce a two-stage post-training strategy to facilitate event-centric alignment, which consists of Supervised Fine-tuning (SFT) and Group Relative Policy Optimization (GRPO) [30] on the EVA-Data. During the GRPO stage, three event-centric rewards are designed to enforce global completeness (via event recall), intra-event attribution (via character matching), and inter-event synchronization (via cross-modal retrieval). Collectively, these post-training objectives enhance the fine-grained alignment of audiovisual events, enabling factually accurate, event-grounded audiovisual caption generation.

Extensive experiments demonstrate that EVA-Cap outperforms existing open-source models across 6 audiovisual benchmarks. For example, on general captioning benchmarks, EVA-Cap achieves a total error rate of 34.6% on video-SALMONN-2 testset [32] and surpasses existing methods by a clear margin of 3.3% on WorldSense [15]. Most importantly, on event-centric evaluation with ChronusAV [7] and AV-SpeakerBench [25], EVA-Cap exhibits substantial advantages with an accuracy of 55.4%, 56.2%, and 40.1% for Visual-to-Audio (V2A) inference, Audio-to-Visual (A2V) grounding, and speaker attribution, respectively. These results underscore the capability of EVA-Cap in generating highly faithful captions with precise temporal alignment and accurate event attribution.

2 Related Work

Audiovisual Video Captioning. Recent advancements in Omni Language Models (OLMs) [32, 22, 59, 52] have significantly improved multimodal perception in complex real-world scenarios, thereby promoting the development of audiovisual captioning [12, 23, 11, 20]. However, existing approaches

remain constrained by the scarcity of high-quality audiovisual caption data, and largely rely on either coarse-grained supervision or focus on limited aspects of the task. For instance, AVoCaDO [4] constructs large-scale audiovisual video-caption pairs through temporal orchestration, but suffers from insufficient factual verification and coarse-grained temporal alignment. UGC-Captioner [49] relies on holistic caption-level supervision, whereas DiaDem [5] and D-ORCA [33] focus narrowly on specific aspects of captions, failing to capture fine-grained event-level correspondences. In contrast, we introduce an event-centric perspective for audiovisual captioning. We construct an event-enhanced dataset with event-centric quality control, and propose an event-centric training strategy that explicitly models event-level alignment, enabling accurate and event-grounded audiovisual caption generation.

Reinforcement Learning for Visual Content Captioning. Reinforcement learning [30, 60, 66] plays an increasingly important role in caption optimization, spanning from visual-only to omni-modal scenarios. CapRL [50] and RubiCap [16] design verifiable rubrics for dense image captioning. OwlCap [67] and VideoCap-R1 [24] improve video captioning by matching key information extracted from visual content. In the audiovisual domain, D-ORCA [33] introduces dialogue-oriented rewards focusing on speaker and ASR consistency, while TimeChat-Captioner [57] optimizes timestamp alignment under structured caption formats. However, these approaches typically focus on limited aspects and struggle to achieve fine-grained audiovisual alignment. In this paper, we develop three event-centric rewards that jointly enforce fine-grained supervision within and across modalities. This approach provides more precise event alignment than prior methods, which either focus on modality-specific objectives [33, 5] or rely on coarse audiovisual synchronization [4, 57].

3 Method

This section introduces the construction of EVA-Cap, which is built upon Qwen-2.5-Omni [51] and optimized for audiovisual video captioning via event-centric alignment. First, we address the scarcity of high-quality data by curating EVA-Data, a 63K-scale audiovisual captioning dataset. The dataset is generated using an automated pipeline with rigorous event-centric quality control (Sec. 3.1). Next, we introduce a two-stage training strategy comprising SFT and GRPO to improve EVA-Cap with event-centric alignment. The training process integrates a suite of event-centric rewards to effectively exploit event-level supervision (Sec. 3.2).

3.1 EVA-Data Construction

3.1.1 Annotation Pipeline

We develop an automatic annotation pipeline to construct audiovisual video–caption pairs with event-level fidelity. As shown in Fig. 2, the pipeline comprises three main stages:

Stage 1: Initialization. We first utilize the open-source audiovisual captioner AVoCaDO [4], which excels in holistic perception, to generate initial audiovisual descriptions for raw videos. This provides a coherent narrative of each video and captures key events and details, covering scene backgrounds, interactive actions, and audio transcriptions. We then filter out samples with repetitive descriptions or abnormal text lengths.

Stage 2: Refinement. We then feed both raw videos and initial descriptions to the advanced omni-model Gemini-2.5-Pro [8] for content refinement and temporal alignment. Guided by raw audiovisual signals, we instruct the model to strengthen the temporal synchronization of audiovisual events and revise event character attribution, while preserving the original natural language style. This yields refined descriptions with precise audiovisual event alignment, achieving fine-grained annotation.

Stage 3: Event-Centric Quality Control. To ensure the quality of the curated data, we introduce an aligned event hit mechanism for the automatic validation of refined descriptions. Specifically, we employ Gemini-2.5-Pro to extract timestamp-grounded audiovisual event tuples directly from raw videos. We then input all event tuples together with refined descriptions to GPT-4.1 [1] to verify whether each event tuple is hit in the caption. An event hit requires both audio and visual events to be entailed with consistent temporal synchronization. This process jointly assesses the accuracy of audiovisual synchronization and event attribution in captions. We discard samples with an event hit ratio below a threshold of 0.8. In this way, we build a large-scale audiovisual caption dataset, EVA-Data, which provides high-quality event-level supervision for event-centric post-training.

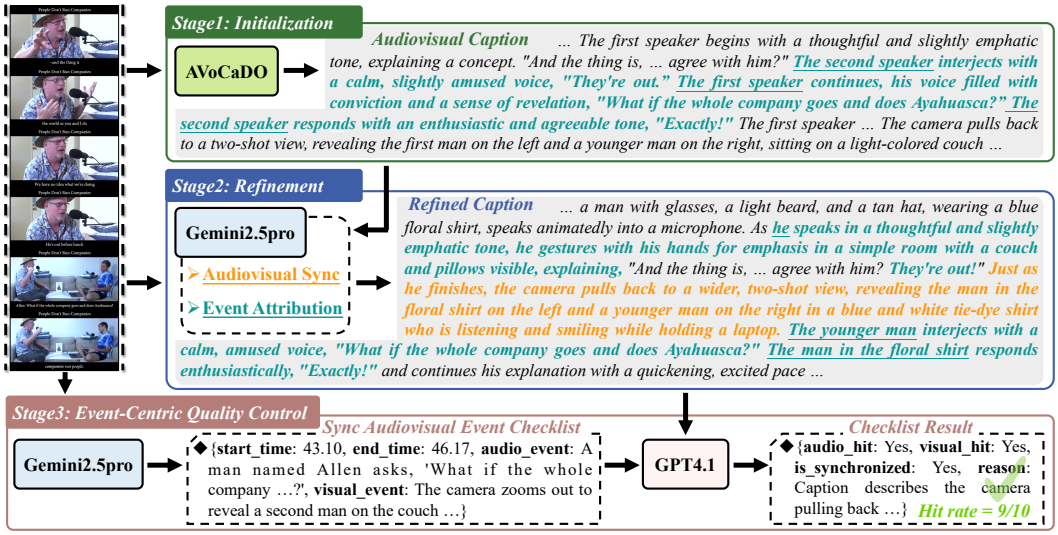
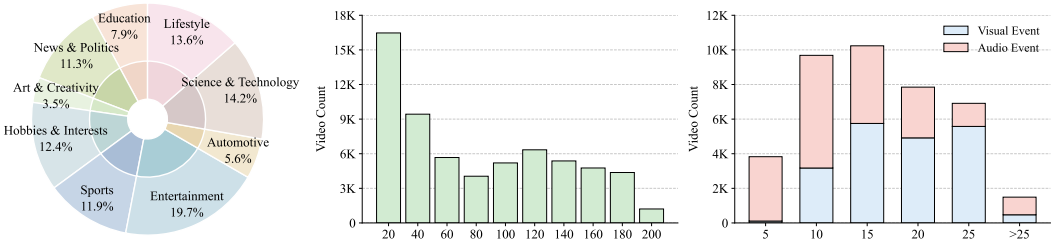


Figure 2: Illustration of the designed data annotation pipeline, comprising three stages: initialization, refinement and event-centric quality control. Raw videos are transformed into audiovisual captions with high-fidelity audiovisual synchronization and event attributions.



(a) Video Category (b) Video Duration (Seconds) (c) Event Count (Audio/Visual)

Figure 3: Statistics of EVA-Data. (a) Video category distribution and (b) Video duration distribution demonstrate the diversity of this dataset, while (c) Event count distribution shows comparable counts of audio and visual events, indicating no modality bias in annotated captions.

3.1.2 Data Statistics

To guarantee both diversity and quality, we collect 100K audiovisual videos (each under 200 seconds in duration) from LLaVA-Video [64], FineVideo [9], and Shot2Story [14]. Through the annotation pipeline, EVA-Data ultimately yields 63K audiovisual video-caption pairs with event-level fidelity. Fig. 3 (b) illustrates the duration distribution of these videos. On average, each video lasts 73.7 seconds, while the corresponding captions average 504.3 words. We further randomly sample 20K videos from EVA-Data and analyze the distribution of audio and visual events in the corresponding captions, as shown in Fig. 3 (c). The results demonstrate that the annotated videos contain rich audiovisual information and the captions are free of modality bias, thus providing effective supervision for event-centric alignment.

3.2 Post-Training of EVA-Cap

Leveraging this dataset, we propose EVA-Cap, an audiovisual captioner optimized via event-centric alignment. We first present our two-stage training pipeline in Sec. 3.2.1, and then introduce the event-centric reward functions dedicated to fine-grained alignment in Sec. 3.2.2.

3.2.1 Training Pipeline

We build EVA-Cap upon the Qwen2.5-Omni Thinker module, which jointly encodes audio and video into a unified, temporally interleaved token sequence to facilitate synchronous audiovisual

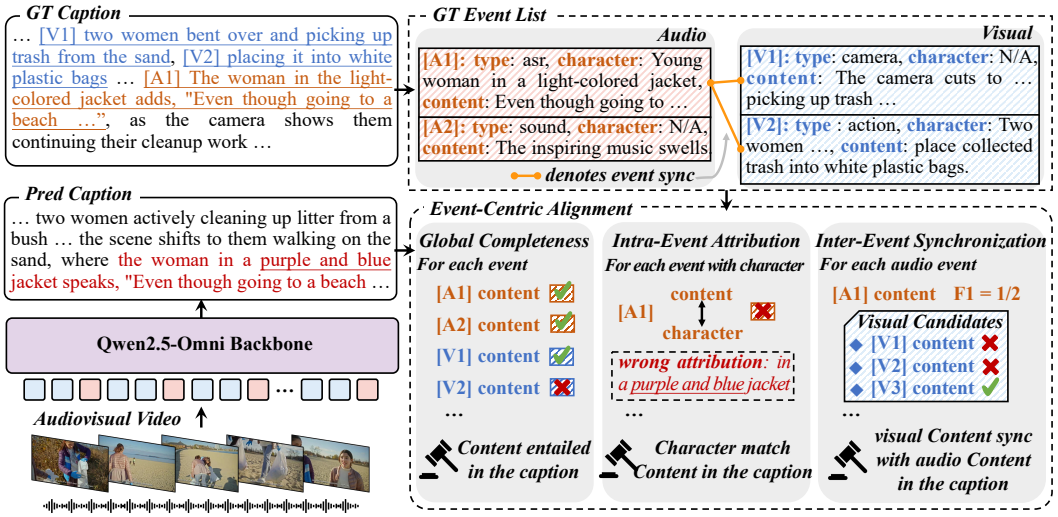


Figure 4: Illustration of the event-centric alignment with GRPO for EVA-Cap. We introduce event-centric rewards during GRPO training, focusing on three key dimensions: Global Completeness (event recall), Intra-Event Attribution (character matching), and Inter-Event Synchronization (cross-modal retrieval between audio and visual events).

understanding. We then adopt a two-stage post-training pipeline consisting of supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO).

In the first phase, we perform SFT on our constructed 63K EVA-Data to empower the model to generate detailed captions that integrate both audio and visual modalities. In the second phase, we further optimize the SFT model with GRPO on 3.2K rollout videos from EVA-Data (details of the rollout process are provided in Appendix A.2.2). To enhance fine-grained alignment in audiovisual captions, we design event-centric rewards covering three dimensions: event completeness, character attribution, and audiovisual synchronization. The detailed SFT and GRPO objectives are provided in Appendix A.1, and the reward design is elaborated in Sec. 3.2.2.

3.2.2 Event-Centric Reward Modeling

Typically, given an audiovisual video $v = (v^{\text{vis}}, v^{\text{aud}})$ and a caption y, conventional preference optimization assigns a holistic reward $\mathcal{R}(y, v)$. However, this paradigm treats the caption as an atomic semantic unit and overlooks its underlying event structure. To address this issue, we represent a caption as a sequence of modality-grounded events:

$$E = (e_1, e_2, \ldots, e_K), \quad e_k = (m_k, \tau_k, c_k, d_k), \quad m_k \in \{\text{vis}, \text{aud}\}, \tag{1}$$

where m_k denotes the modality, τ_k denotes the event type, c_k denotes the character, and d_k denotes the event content.

Based on this event-centric representation, we further define the reference structures to support the event-centric reward calculation. For a reference caption $y^\star$, we denote its reference event set as $\mathcal{E}^\star = \mathcal{E}^\star_{\text{aud}} \cup \mathcal{E}^\star_{\text{vis}}$, where $\mathcal{E}^\star_m = \{e^m_i\}_{i=1}^{N_m}$ is the set of modality-specific reference events for $m \in \{\text{aud}, \text{vis}\}$. This event set is used to evaluate event coverage and character attribution. In addition, we denote by $\mathcal{Q}^\star = \{q^\star_i\}_{i=1}^{M}$ the set of audio-to-visual synchronization queries, where each query consists of an audio event, a set of candidate visual events, and the reference synchronized visual indices. This query set is used to evaluate cross-modal temporal synchronization.

Based on the above event formulation and definitions, we devise three event-centric reward components to achieve fine-grained alignment (as shown in Fig. 4):

Global Event Completeness. Global event completeness measures whether the generated caption covers the reference audio and visual events that should be described in the video. Given a reference event set $\mathcal{E}^\star = \mathcal{E}^\star_{\text{aud}} \cup \mathcal{E}^\star_{\text{vis}}$, where $\mathcal{E}^\star_m = \{e^m_i\}_{i=1}^{N_m}$ denotes the modality-specific events for $m \in$

$\{\text{aud}, \text{vis}\}$, we define an LLM-evaluated coverage indicator:

$$\mathbb{I}^m_{\text{cov}}(e^m_i, y) = \begin{cases} 1, & \text{if Covers}(y, e^m_i), \\ 0, & \text{otherwise,} \end{cases} \quad m \in \{\text{aud}, \text{vis}\}. \tag{2}$$

Here, $\text{Covers}(y, e^m_i)$ indicates that the event content of e^m_i is semantically supported by the generated caption y. The global event completeness reward is defined as event recall over all reference audio and visual events:

$$\mathcal{R}_{\text{comp}}(y, \mathcal{E}^\star) = \frac{\sum_{m\in\{\text{aud},\text{vis}\}} \sum_{i=1}^{N_m} \mathbb{I}^m_{\text{cov}}(e^m_i, y)}{\sum_{m\in\{\text{aud},\text{vis}\}} N_m}. \tag{3}$$

Intra-Event Attribution. Event completeness alone does not guarantee that a caption assigns events to the correct entities. For example, a caption may mention the correct speech content but attribute it to the wrong speaker. To address this issue, we introduce an intra-event attribution reward that evaluates whether each character-bearing reference event is both covered and associated with the correct character. Given the reference event set $\mathcal{E}^\star = \mathcal{E}^\star_{\text{aud}} \cup \mathcal{E}^\star_{\text{vis}}$, where $\mathcal{E}^\star_m = \{e^m_i\}_{i=1}^{N_m}$ denotes the modality-specific events for $m \in \{\text{aud}, \text{vis}\}$, we define an LLM-evaluated attribution indicator for each valid event with an explicit character:

$$\mathbb{I}^m_{\text{attr}}(e^m_i, y, y^\star) = \begin{cases} 1, & \text{if Matches}(y, y^\star, e^m_i), \\ 0, & \text{otherwise,} \end{cases} \quad m \in \{\text{aud}, \text{vis}\}. \tag{4}$$

Here, $\text{Matches}(y, y^\star, e^m_i)$ indicates that the generated caption y and the reference caption $y^\star$ attribute the event content d^m_i to the same underlying character. The attribution reward is computed as

$$\mathcal{R}_{\text{attr}}(y, y^\star, \mathcal{E}^\star) = \frac{\sum_{m\in\{\text{aud},\text{vis}\}} \sum_{i=1}^{N_m} \delta^m_{i,\text{char}} \mathbb{I}^m_{\text{cov}}(e^m_i, y) \mathbb{I}^m_{\text{attr}}(e^m_i, y, y^\star)}{\sum_{m\in\{\text{aud},\text{vis}\}} \sum_{i=1}^{N_m} \delta^m_{i,\text{char}}}, \tag{5}$$

where $\delta^m_{i,\text{char}} = \mathbb{I}[c^m_i \neq \varnothing]$ indicates whether the event has an explicit character. Character-free events are not counted in this reward.

Inter-Event Synchronization. To ensure accurate temporal correspondence between audio and visual events, we model cross-modal alignment explicitly by introducing an inter-event synchronization reward, which is formulated as audio-to-visual synchronization retrieval. Given a set of audio synchronization queries $\mathcal{Q}^\star = \{q^\star_i\}_{i=1}^M$, each query consists of an audio event e^{aud}_i, a candidate visual event set $\mathcal{C}^{\text{vis}}_i = \{v_{i,j}\}_{j=1}^{L_i}$, and the reference synchronized visual index set $\mathcal{G}_i \subseteq \{1, \ldots, L_i\}$. For each query, an LLM evaluator selects the visual candidates that are temporally synchronized with the audio event according to the generated caption y:

$$\hat{\mathcal{G}}_i = \text{RetrieveSync}\left(y, e^{\text{aud}}_i, \mathcal{C}^{\text{vis}}_i\right), \tag{6}$$

where $\hat{\mathcal{G}}_i \subseteq \{1, \ldots, L_i\}$ denotes the predicted synchronized visual index set. If no synchronized visual event is found, RetrieveSync returns an empty set. We then compute the set-level F1 score:

$$\text{Prec}_i = \frac{|\hat{\mathcal{G}}_i \cap \mathcal{G}_i|}{|\hat{\mathcal{G}}_i|}, \quad \text{Rec}_i = \frac{|\hat{\mathcal{G}}_i \cap \mathcal{G}_i|}{|\mathcal{G}_i|}, \quad F_i = \frac{2\text{Prec}_i\text{Rec}_i}{\text{Prec}_i + \text{Rec}_i}. \tag{7}$$

When both $\hat{\mathcal{G}}_i$ and $\mathcal{G}_i$ are empty, we set $F_i = 1$, when only one of them is empty, or their intersection is empty, we set $F_i = 0$. The synchronization reward is then defined as

$$\mathcal{R}_{\text{sync}}(y, \mathcal{Q}^\star) = \frac{1}{M} \sum_{i=1}^{M} F_i. \tag{8}$$

We therefore formulate our overall reward $\mathcal{R}_{\text{EVA}}$ with three complementary event-centric terms:

$$\mathcal{R}_{\text{EVA}}(y, y^\star) = \lambda_{\text{comp}}\mathcal{R}_{\text{comp}}(y, \mathcal{E}^\star) + \lambda_{\text{attr}}\mathcal{R}_{\text{attr}}(y, y^\star, \mathcal{E}^\star) + \lambda_{\text{sync}}\mathcal{R}_{\text{sync}}(y, \mathcal{Q}^\star). \tag{9}$$

Here, λ_{comp}, λ_{attr}, and λ_{sync} are balancing coefficients. For all reward terms, if the corresponding valid evaluation set is empty, i.e., $\sum_m N_m = 0$, $\sum_m \sum_i \delta^m_{i,\text{char}} = 0$, or $M = 0$, we set the corresponding reward value to 0.

During training, we adopt this reward with two rounds of GRPO to achieve progressive alignment. Specifically, in the first round, we set λ_{comp} and λ_{sync} to 1 while setting λ_{attr} to 0. In the second round, we set all three coefficients to 1. The process transitions from global coarse-grained event alignment and audiovisual synchronization to fine-grained character-level attribution.

4 Experiments

4.1 Experimental Setup

Implementation Details: We adopt Qwen2.5-Omni-7B [51] as our backbone model. During training, the vision and audio encoders are kept frozen, while only the adapters and LLM parameters are updated. Videos are sampled at 2 FPS, with a maximum per-frame pixels of $512 \times 28 \times 28$ and total pixels of $30 \times 512 \times 28 \times 28$. And the maximum generation length is set to 2048. For SFT, we train the model for two epochs with a batch size of 128 and a learning rate of 1×10^{-5}. For GRPO, we conduct training for two epochs with a batch size of 16, a learning rate of 2×10^{-6}, a KL penalty of 0.04, and a group size of 8, following a progressive reward weighting strategy. We employ Qwen3.5-27B [36] as the reward judge. All training is conducted on 32 NVIDIA H20 GPUs.

Benchmarks: For general audiovisual captioning, we evaluate EVA-Cap under two settings: direct quality assessment (video-SALMONN-2 testset [32] and UGC-VideoCap [49]) and QA-cascaded evaluation (Daily-Omni [68] and WorldSense [15]). The former directly evaluates the overall quality of generated captions through dedicated scoring mechanisms, while for the latter, we first generate descriptive captions with our model and then adopt Gemini-2.5-Pro [8] to answer QA questions solely based on the produced captions.

To assess the capabilities in event-centric alignment, we further conduct evaluations on subsets of ChronusAV [7] and AV-SpeakerBench [25]. For ChronusAV, we select 1,066 videos under 200 seconds to assess audiovisual synchronization via video-to-audio (V2A) and audio-to-video (A2V) grounding. Evaluation follows a two-stage protocol: Gemini-2.5-Pro first generates answers for the V2A and A2V grounding queries conditioned on the model-generated captions, and Seed-1.6 [34] subsequently scores the outputs to compute VDCscore [2] and accuracy. For AV-SpeakerBench, we adopt 422 QA samples from its Speaker Recognition subset to assess audio event attribution, using the same QA-cascaded evaluation paradigm with Gemini-2.5-Pro.

Table 1: Performance comparison on general audiovisual video captioning benchmarks. Following previous works [4, 5], we adopt GPT-4.1 [1] as the judge model for the video-SALMONN-2 testset, and GPT-4o [19] for UGC-VideoCap evaluation.

Model	**Size**	**video-SALMONN-2 testset**			**UGC-VideoCap**			
		Miss.↓	Hall.↓	Total↓	Audio↑	Visual↑	Detail↑	Avg.↑
Gemini-2.5-Pro [8]	-	18.1	13.3	31.3	69.5	74.7	73.7	72.6
Gemini-2.5-Flash [8]	-	19.3	13.9	33.3	69.1	75.8	74.0	73.0
HumanOmniV2 [55]	7B	49.2	**12.3**	61.6	45.6	66.3	59.5	57.1
ARC-Hunyuan-Video [11]	7B	45.7	12.5	58.2	52.7	56.0	55.8	54.8
Qwen2.5-Omni [51]	7B	41.7	15.4	57.1	46.9	66.1	60.0	57.7
MiniCPM-o-2.6 [58]	8B	42.2	14.3	56.5	38.6	68.5	57.7	54.9
UGC-VideoCaptioner [49]	3B	31.6	17.0	48.6	61.4	58.4	57.5	59.1
Qwen3-Omni-Instruct [52]	30B-A3B	32.0	13.6	45.6	67.5	74.8	72.3	71.5
Qwen3-Omni-Captioner [52]	30B-A3B	31.0	16.6	47.6	69.0	75.5	72.3	72.3
video-SALMONN-2 [32]	7B	21.2	17.6	38.8	61.8	71.4	68.5	67.2
AVoCaDO [4]	7B	21.1	16.2	37.3	73.0	74.6	71.8	73.2
DiaDem [5]	7B	21.1	15.5	36.5	75.4	76.8	74.6	75.6
EVA-Cap	7B	**17.6**	17.0	**34.6**	**76.9**	**79.0**	**74.7**	**76.9**

4.2 Main Results

Performance on direct quality assessment. We evaluate EVA-Cap on two representative general audiovisual captioning benchmarks. Those benchmarks assess the overall captioning performance from multiple dimensions, including video content omission, hallucination, and multi-aspect scoring. The comparison with existing audiovisual models is presented in Tab. 1. EVA-Cap achieves state-of-the-art or competitive results across various metrics, *e.g.*, a total error rate of 34.6% on the video-SALMONN-2 testset [32] and an average score of 76.9 on UGC-VideoCap [49], surpassing both unified OLMs, *e.g.*, Qwen3-Omni-Instruct [52] and task-specific captioners, *e.g.*, DiaDem [5].

On the video-SALMONN-2 test set, EVA-Cap demonstrates the lowest miss rate of 17.6%. While some baselines achieve lower hallucination rates, they often do so at the cost of substantial event omissions, e.g., HumanOmniV2 maintains a 12.3% hallucination rate at the cost of a 49.2% miss rate. This indicates OLMs may omit many valid yet less salient events without event-centric training. This demonstrates the strong general audiovisual captioning capability of our method.

Performance on QA-cascaded evaluation. To further evaluate the information completeness and accuracy of the generated captions, we conduct reasoning-based QA evaluation under the cascaded paradigm described above. Specifically, we prompt Gemini-2.5-Pro [8] to answer questions based on the model-generated captions, and the resulting QA accuracy reflects the overall quality of audiovisual captions. As shown in Tab. 2, EVA-Cap outperforms existing open-source models by a clear margin, with gains of +3.3% on Daily-Omni [68] and +3.3% on WorldSense [15].

Table 2: Performance comparison on QA-cascaded benchmarks, with Gemini-2.5-Pro serving as the judge.

Model	Size	Daily-Omni	World-Sense
Gemini-2.5-Pro [8]	-	60.2	33.8
Gemini-2.5-Flash [8]	-	55.3	31.0
HumanOmniV2 [55]	7B	8.2	6.6
ARC-Hunyuan-Video [11]	7B	8.6	8.7
MiniCPM-o-2.6 [58]	8B	9.8	7.2
Qwen2.5-Omni [51]	7B	12.4	8.6
UGC-VideoCaptioner [49]	3B	17.0	11.2
Qwen3-Omni-Instruct [52]	30B-A3B	17.5	12.7
Qwen3-Omni-Captioner [52]	30B-A3B	27.2	14.1
video-SALMONN-2 [32]	7B	29.9	18.2
AVoCaDO [4]	7B	50.1	25.7
TimeChat-Captioner [57]	7B	52.8	22.6
EVA-Cap	7B	**56.1**	**29.0**

Performance on event-centric alignment. Finally, we evaluate event-centric alignment on ChronusAV [7] and AV-SpeakerBench [25]. For ChronusAV, we adopt METEOR to measure textual similarity, and further compute accuracy and VDCscore following the protocol of the VDC benchmark [2]. For AV-SpeakerBench, we take the accuracy of multiple-choice questions as the evaluation metric. For comparison, we include three advanced open-source models that achieve competitive results on general audiovisual captioning benchmarks, *i.e.*, TimeChat-Captioner [57], AVoCaDO [4], and DiaDem [5].

As shown in Tab. 3, EVA-Cap significantly outperforms the aforementioned models across all metrics. This not only validates the capability of our model to attribute events to the corresponding characters, but also highlights its strength in achieving fine-grained, precise audiovisual event synchronization.

Table 3: Performance comparison on event-centric evaluation.

Model	Size	V2A			A2V			AV-Speaker
		M	Acc	Score	M	Acc	Score	Acc
Gemini-2.5-Pro [8]	-	2.4	61.2	2.4	1.2	61.0	2.4	49.1
TimeChat [57]	7B	1.2	31.6	1.2	0.7	30.1	1.2	23.7
AVoCaDO [4]	7B	1.8	40.9	1.6	1.1	44.9	1.8	26.8
DiaDem [5]	7B	1.8	45.9	1.7	1.1	47.7	1.9	32.7
EVA-Cap	7B	**2.4**	**55.4**	**2.2**	**1.2**	**56.2**	**2.2**	**40.1**

Table 4: Ablation study on the post-training stages and event-centric reward mechanisms of EVA-Cap. The two-stage GRPO training setting we adopted is marked in light gray. All GRPO stage1 settings are derived from the SFT model, whereas GRPO stage 2 settings build upon the marked stage1.

Setting	Reward			V2A			A2V			AV-Speaker	Daily-Omni
	$\mathcal{R}_{\text{comp}}$	$\mathcal{R}_{\text{sync}}$	$\mathcal{R}_{\text{attr}}$	M	Acc	Score	M	Acc	Score	Acc	Acc
Baseline	-	-	-	0.1	8.4	0.2	0.3	11.9	0.5	3.1	12.4
SFT	-	-	-	2.0	44.8	1.7	1.0	46.1	1.8	31.8	50.1
SFT → GRPO *stage1*	✓	-	-	2.2	53.3	2.0	1.2	52.9	2.0	33.9	53.8
SFT → GRPO *stage1*	✓	✓	-	2.3	53.8	2.1	1.2	55.0	2.1	37.0	55.0
SFT → GRPO *stage1*	✓	✓	✓	2.1	47.9	1.8	1.1	50.8	1.9	35.6	52.9
SFT → GRPO *stage2*	✓	✓	-	2.2	**56.5**	**2.2**	**1.3**	55.5	**2.2**	37.0	55.1
SFT → GRPO *stage2*	✓	✓	✓	**2.4**	55.4	**2.2**	1.2	**56.2**	**2.2**	**40.1**	**56.1**



AVoCaDO

The audio opens with a soft, emotional, … A close-up shot focuses on a small, brown, animated puppy … **A young boy's voice, filled with concern and empathy, begins to speak. "I know. You saw me and took pity on me." The camera pulls back to a wider shot, revealing the puppy sitting in the middle of a crosswalk. A young boy with brown hair, wearing a white t-shirt and jeans, stands in front of the puppy**, … The scene shifts to a daytime setting. The puppy is now sitting on the sidewalk, … The boy's voice becomes firm and reassuring, his tone filled with admiration and conviction. **"Not at all … kept going." The view returns to the nighttime street scene. The boy is standing in front of the truck again** …

EVA-Cap

The audio begins with a soft, melancholic, … a close-up shot of a small, brown animated puppy … The scene then cuts to a wider shot where the puppy is now sitting on the crosswalk, looking up at a young boy with brown hair wearing a white t-shirt and blue jeans who stands before him … **The camera angle shifts to show the boy, now wearing a red and blue vest, kneeling on the sidewalk next to the puppy, which sits looking up at him with a worried expression against a brick wall and other people walking in the background. Simultaneously, the puppy speaks in a soft, hesitant voice filled with vulnerability, "I know. You saw me and took pity on me."** The boy responds in a firm but kind tone, **"Not at all," and as he continues speaking, the camera cuts back to a tight close-up of the puppy's face, showing its large, sad eyes.** The boy's voice becomes warm and encouraging, stating, **"What I saw was a brave, heroic pup," while the shot returns to a medium view of both characters** …

Figure 5: Case analysis of EVA-Cap and AVoCaDO, with features relevant to **audiovisual synchronization** in bold, and accurate event attribution underlined. Incorrect event-centric alignments are marked in red, and correct ones in green.

4.3 Ablation Studies

In Tab. 4, we conduct ablation experiments to study the effects of our post-training pipeline and reward suite. First, the baseline Qwen2.5-Omni suffers from inherent modality bias [29], where audio content is frequently omitted in its generated captions, resulting in poor performance across all tasks. After SFT on our constructed EVA-Data, the model achieves considerable improvements in both general audiovisual captioning and event alignment, e.g., the V2A and A2V accuracy increases from 8.4% to 44.8% and from 11.9% to 46.1%, respectively.

We further investigate the effects of three event-centric rewards during the GRPO phase. The global event completeness reward $\mathcal{R}_{\text{comp}}$ penalizes event omission, thereby boosting model performance across all dimensions (+8.5%, +2.1%, and +3.7% accuracy on V2A, AV-Speaker, and Daily-Omni). Furthermore, integrating the inter-event synchronization reward $\mathcal{R}_{\text{sync}}$ and intra-event attribution reward $\mathcal{R}_{\text{attr}}$ yields consistent gains on corresponding tasks (+2.1% on A2V and +3.1% on AV-Speaker in terms of accuracy), while also improving the overall captioning performance steadily. Collectively, these three rewards enable precise alignment in both event content and event correspondence, facilitating factually accurate, event-grounded audiovisual video captioning.

To further dissect our two-stage GRPO design, we conduct additional ablations. Introducing attribution reward $\mathcal{R}_{\text{attr}}$ as early as the first round not only compromises general performance but also degrades event-centric alignment. We hypothesize that injecting both $\mathcal{R}_{\text{sync}}$ and $\mathcal{R}_{\text{attr}}$ prematurely leads to optimization conflicts between inter-event relations and intra-event accuracy. Conversely, a progressive strategy facilitates a gradual alignment process, from global coarse-grained structures to fine-grained character-level details. Moreover, comparisons against a baseline that sustains only $\mathcal{R}_{\text{comp}}$ and $\mathcal{R}_{\text{sync}}$ in the second round demonstrate that the sustained improvements stem from effective reward design rather than mere data scaling.

4.4 Visualizations

We present a caption generated by EVA-Cap in Fig. 5 for comparison with AVoCaDO [4], highlighting the superior event alignment capabilities (*i.e.*, audiovisual synchronization and event attribution) of our model. For instance, the audio ASR event "Puppy: I know. You saw me and took pity on me" occurs when characters converse on the sidewalk. AVoCaDO incorrectly synchronizes it with the adjacent scene in the middle of a crosswalk and misattributes the event to the boy. In contrast, our model precisely aligns these audiovisual events and accurately identifies the corresponding speaker. This example demonstrates that our model can produce high-fidelity captions in complex audiovisual

scenarios involving multi-view transitions and multi-object interactions, and further underscores the efficacy of the event-centric alignment mechanism for long-form audiovisual captioning.

5 Conclusion

This paper presents EVA-Cap, an audiovisual video captioning model with a training framework that transforms holistic caption optimization into event-centric alignment. To support event-level supervision, we first constructed EVA-Data with event-centric quality control and adopted a two-stage post-training strategy with SFT and GRPO. Then, by decomposing captions into modality-grounded events, EVA-Cap elaborates fine-grained GRPO rewards tailored to event completeness, character attribution, and audiovisual synchronization. Extensive experiments demonstrate that SFT on EVA-Data significantly improves OLMs to internalize event-centric knowledge. Furthermore, the subsequent GRPO optimization improves event-centric alignment while maintaining strong general captioning performance. These results suggest that event-centric alignment offers an effective path toward more faithful and temporally coherent audiovisual video captioning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472098) and the Science and Technology Commission of Shanghai Municipality (No. 25511106100).

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] W. Chai, E. Song, Y. Du, C. Meng, V. Madhavan, O. Bar-Tal, J.-N. Hwang, S. Xie, and C. D. Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *ICLR*, 2025.

[3] L. Che, B. Liu, D. Li, X. Li, Z. Wu, C. Li, M. Guo, C. Mao, V. Pavlovic, and R. Krishna. Lvmh: A meta-optimization harness for long-video agent memory. *arXiv preprint arXiv:2610.23904*, 2026.

[4] X. Chen, Y. Ding, W. Lin, J. Hua, L. Yao, Y. Shi, B. Li, Y. Zhang, Q. Liu, P. Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *ICLR*, 2026.

[5] X. Chen, W. Lin, J. Hua, L. Yao, Y. Ding, B. Li, B. Zeng, Y. Shi, Q. Liu, Y. Zhang, et al. Diadem: Advancing dia logue de scriptions in audiovisual video captioning for multi m odal large language models. In *ECCV*, 2026.

[6] X. Chen, J. Tang, Y. Ding, Y. Jia, B. Li, B. Zeng, Y. Shi, S. Li, Y. Ji, Q. Liu, et al. Capricorn-1k: A comprehensive benchmark for video captioning and subject referential consistency across temporal scales. In *EMNLP*, 2026.

[7] Y. Chen, Y. Wu, K. Guan, Y. Ren, Y. Wang, R. Song, and L. Ru. Chronusomni: Improving time awareness of omni large language models. In *ECCV*, 2026.

[8] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[9] M. Farré, A. Marafioti, L. Tunstall, L. Von Werra, and T. Wolf. Finevideo, 2024.

[10] K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, J. Wu, X. Zhang, B. Wang, and X. Yue. Video-r1: Reinforcing video reasoning in mllms. In *NeurIPS*, 2025.

[11] Y. Ge, Y. Ge, C. Li, T. Wang, J. Pu, Y. Li, L. Qiu, J. Ma, L. Duan, X. Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

[12] M. Guo, Q. Jiao, X. Liu, J. Pang, R. Yang, N. Xu, L. Che, H. Xue, Z. Wang, X. Cai, et al. Seeing, maintaining, and learning: The evolution of multimodal agent memory. *arXiv preprint arXiv:2610.19366*, 2026.

[13] Y. Guo, J. Liu, M. Li, Q. Liu, X. Chen, and X. Tang. Trace: Temporal grounding video llm via causal event modeling. In *ICLR*, 2025.

[14] M. Han, L. Yang, X. Chang, L. Yao, and H. Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *ICLR*, 2025.

[15] J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *ICLR*, 2026.

[16] T.-H. Huang, S. Salekin, J. Movellan, F. Sala, and M. Bilkhu. Rubicap: Rubric-guided reinforcement learning for dense image captioning. *arXiv preprint arXiv:2603.09160*, 2026.

[17] Y. Huang, S. Tu, X. Yan, Z. Yang, J. Han, H. Xu, K. Pan, Y.-G. Jiang, and Z. Wu. Va-judger: Reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026.

[18] Y. Huang, S. Tu, X. Yan, J. Zhan, Z. Yang, Z. Xing, H. Zhang, T. Zhang, Y.-G. Jiang, and Z. Wu. Agentic visual generation: From generative models to agentic control. *arXiv preprint arXiv:2609.06758*, 2026.

[19] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[20] S. Kumar, P. Seetharaman, K. Chen, O. Nieto, J. Su, Z. Wang, R. Kumar, D. Manocha, N. J. Bryan, Z. Jin, et al. Tac: Timestamped audio captioning. *arXiv preprint arXiv:2602.15766*, 2026.

[21] Y. Li, H. Zhang, M.-H. Guo, W. Gao, S. Jia, S. Jiao, Q. Hou, and M.-M. Cheng. Towards universal video mllms with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026.

[22] Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

[23] Z. Ma, R. Xu, Z. Xing, Y. Chu, Y. Wang, J. He, J. Xu, P.-A. Heng, K. Yu, J. Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *ICLR*, 2026.

[24] D. Meng, R. Huang, Z. Dai, X. Li, Y. Xu, J. Zhang, Z. Huang, M. Zhang, L. Zhang, Y. Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.

[25] L. T. P. Nguyen, Z. Yu, S. L. Y. Hang, S. An, J. Lee, Y. Ban, S. Chung, T.-H. Nguyen, J. Maeng, S. Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. In *CVPR Findings*, 2026.

[26] W. Peng, L. Meng, Y. Chen, Y. Xie, Y. Liu, T. Gui, H. Xu, X. Qiu, Z. Wu, and Y.-G. Jiang. Inst-it: Boosting instance understanding via explicit visual prompt instruction tuning. In *NeurIPS*, 2025.

[27] J. Pu, Y. Chen, T. Wang, and Y. Shan. Omniscript: Towards audio-visual script generation for long-form cinematic video. In *ECCV*, 2026.

[28] K. Qiu, M. Han, K. Liu, Y. Liu, J. Han, Y. Jiang, X. Wu, S. Wang, L. Zhang, and D. Yang. Omnifysics-captioner technical report: Grounding omni-modal understanding in the physical world for better captioning. *arXiv preprint arXiv:2609.31714*, 2026.

[29] R. Selvakumar, K. Jayakumar, S. Sakshi, S. Ghosh, R. Gao, and D. Manocha. Do audio-visual large language models really see and hear? In *CVPR Findings*, 2026.

[30] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[31] J. Shi, J. Wang, Z. You, B. He, and Z. Wu. Videoloom: A video large language model for joint spatial-temporal understanding. In *ICML*, 2026.

[32] C. Tang, Y. Li, Y. Yang, J. Zhuang, G. Sun, W. Li, Z. Ma, and C. Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

[33] C. Tang, T. Wang, F. Rao, J. Lyu, and C. Zhang. D-orca: Dialogue-centric optimization for robust audio-visual captioning. *arXiv preprint arXiv:2602.07960*, 2026.

[34] B. S. Team. Introduction to techniques used in seed1. 6, 2025.

[35] G. Team, S. E. Abd, V. Aggarwal, R. Algayres, A. Andreev, O. Bachem, I. Ballantyne, C. Brick, V. Cărbune, M. Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

[36] Q. Team. Qwen3.5: Accelerating productivity with native multimodal agents, 2026.

[37] R. Tian, Q. Dai, H. Hu, and Z. Wu. Tinymem: Condensing multimodal memory for long-form video action detection. *TPAMI*, 2026.

[38] S. Tu, Q. Dai, Z. Wu, Z.-Q. Cheng, H. Hu, and Y.-G. Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023.

[39] S. Tu, Y. Pan, Y. Huang, X. Han, Z. Xing, Q. Dai, C. Luo, Z. Wu, and Y.-G. Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. In *NeurIPS*, 2026.

[40] S. Tu, Q. Tian, Y. Huang, Y. Wu, X. Han, K. Pan, W. Kong, J. Xiong, J.-W. Zhang, Z. Wu, et al. Prism: Dynamic sparse attention for native 2k joint video-audio generation model training. *arXiv preprint arXiv:2610.05416*, 2026.

[41] S. Tu, Q. Tian, Z. Yang, Y. Wu, X. Han, W. Kong, J. Xiong, J.-W. Zhang, Z. Zhong, L. Bo, et al. Baton: Explicit semantic blueprints for joint video-audio generation. In *NeurIPS*, 2026.

[42] J. Wang, D. Chen, C. Luo, X. Dai, L. Yuan, Z. Wu, and Y.-G. Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023.

[43] J. Wang, D. Chen, C. Luo, B. He, L. Yuan, Z. Wu, and Y.-G. Jiang. Omnivid: A generative framework for universal video understanding. In *CVPR*, 2024.

[44] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022.

[45] J. Wang, L. Yuan, Y. Zhang, and H. Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

[46] W. Wang and Y. Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. In *NeurIPS*, 2025.

[47] Y. Wang, J. Wang, J. Tang, Y. Zhang, Z. Cao, H. Bian, Z. Zhang, W. Luo, Z. Pan, Z. Dong, J. Liu, and Z. Zhang. Avscap: Orchestrating audio-visual synergy for omni-modal video captioning. In *EMNLP*, 2026.

[48] M. Wu, K. Feng, B. Li, K. Gong, Z. Yin, and X. Yue. Avcap: Reinforcing audio-video joint caption with detail-aware reward. *arXiv preprint arXiv:2608.06930*, 2026.

[49] P. Wu, Y. Liu, Z. Zhu, E. Zhou, and J. Shen. Ugc-videocaptioner: An omni ugc video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.

[50] L. Xing, X. Dong, Y. Zang, Y. Cao, J. Liang, Q. Huang, J. Wang, F. Wu, and D. Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *ICLR*, 2026.

[51] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[52] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[53] Y. Xu, X. Li, Y. Yang, D. Meng, R. Huang, and L. Wang. Carebench: A fine-grained benchmark for video captioning and retrieval. In *ICLR*, 2026.

[54] Y. Xu, Z. Wang, Z. Wang, J. Zhang, Y. Yang, D. Meng, Y. Zhang, P. Wan, and L. Wang. Percepcap: Video captioner with structured spatio-temporal perception. In *NeurIPS*, 2026.

[55] Q. Yang, S. Yao, W. Chen, S. Fu, D. Bai, J. Zhao, B. Sun, B. Yin, X. Wei, and J. Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

[56] Z. Yang, J. Tao, R. Chen, Z. Yang, Y. Yuan, X. Zhao, Auden, K. Wang, S. Shao, B. Wang, S. Yves, and Q. Lu. Omnicapbench: A deep-structured evaluation framework for fine-grained audio-visual captioning. In *NeurIPS*, 2026.

[57] L. Yao, Y. Wei, Y. Zhang, L. Li, X. Chen, F. Song, Z. Wang, K. Ouyang, Y. Liu, L. Kong, et al. Timechat-captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. In *ICML*, 2026.

[58] Y. Yao, T. Yu, C. Wang, J. Cui, B. Xu, H. Zhu, T. Cai, F. Huang, T. Wang, W. Ma, et al. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2025.

[59] H. Ye, C.-H. H. Yang, A. Goel, W. Huang, L. Zhu, Y. Su, S. Lin, A.-C. Cheng, Z. Wan, J. Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *ICLR*, 2026.

[60] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. In *NeurIPS*, 2025.

[61] L. Yuan, J. Wang, H. Sun, Y. Zhang, and Y. Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

[62] Y. Yuan, H. Zhang, W. Li, Z. Cheng, B. Zhang, L. Li, X. Li, D. Zhao, W. Zhang, Y. Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *CVPR*, 2025.

[63] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, 2024.

[64] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. Llava-video: Video instruction tuning with synthetic data. *TMLR*, 2025.

[65] C. Zhao, J. Ma, Q. Huang, T. Fan, H. Li, Z. Kang, X. Wei, J. Yang, and Y. Tai. Temporal and cross-modal alignment for enhanced audiovisual video captioning. In *ECCV*, 2026.

[66] C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[67] C. Zhong, Q. Hou, Z. Zhou, Y. Zhang, S. Hao, H. Lu, H. Tang, and X. Bai. Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. In *AAAI*, 2026.

[68] Z. Zhou, R. Wang, Z. Wu, and Y.-G. Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Technical appendices and supplementary material

A.1 Details about the Training Objectives

A.1.1 Formulation of Supervised Fine-Tuning

In the first stage, we perform supervised fine-tuning (SFT) on EVA-Data to enable the model to generate long-form audiovisual captions that integrate both audio and visual information. Given a video-instruction pair (v, q) and its ground-truth caption $y = (y_1, \ldots, y_T)$, the model is optimized with the standard next-token prediction objective:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(v,q,y)\sim\mathcal{D}_{\mathrm{SFT}}} \left[\log \pi_\theta(y \mid v, q)\right], \tag{10}$$

where $\mathcal{D}_{\mathrm{SFT}}$ denotes the supervised training set, and θ represents the model parameters. This stage provides the initialization for the subsequent GRPO optimization.

A.1.2 Formulation of Group Relative Policy Optimization

In the second stage, we optimize the SFT model with GRPO. For each audiovisual instruction input q, where q contains the video v and the textual instruction, the old policy $\pi_{\theta_{\mathrm{old}}}$ samples a group of outputs $\{o_1, o_2, \ldots, o_G\}$. Each output is evaluated by the event-centric reward function to obtain $R_i = R_{\mathrm{EVA}}(o_i, y^\star), \quad i = 1, \ldots, G$, where $y^\star$ denotes the reference caption and its associated reference event structures. The advantage of the i-th response is computed by normalizing its reward within the sampled group: $\hat{A}_i = \frac{R_i - \mathrm{mean}(\{R_1,\ldots,R_G\})}{\mathrm{std}(\{R_1,\ldots,R_G\})}$. The policy is optimized with the following clipped surrogate objective:

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{q\sim\mathcal{P}(q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \\ \min\left(r_{i,t}(\theta)\hat{A}_i, \mathrm{clip}(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\right) - \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) \Bigg],\end{aligned} \tag{11}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{12}$$

Here, $D_{\mathrm{KL}}(\cdot)$ denotes the KL divergence to the reference policy π_{ref}, ε is the clipping range, and β controls the KL penalty.

A.2 More Implementation Details

A.2.1 Details about the Construction of Reference Structures

We filter 6.4K videos from EVA-Data, all of which achieve a 100% hit ratio in Stage 3 of the annotation pipeline, ensuring highly accurate event-level supervision signals. To facilitate efficient reward computation for both GRPO training and rollout, we construct reference structures for these 6.4K selected videos. Specifically, this involves extracting audiovisual event units $\mathcal{E}^\star$ from the ground-truth captions $y^\star$ and formulating audio-to-visual synchronization queries $\mathcal{Q}^\star$.

We employ the reward judge LLM, Qwen3.5-27B [36], prompting it to sequentially extract audio events and visual events from the given caption to form a unified event list $\mathcal{E}^\star$. Each extracted audio event serves as a separate query. We input the complete ground-truth caption along with all the extracted visual events into the LLM, prompting it to select the temporally synchronized visual events corresponding to the current audio event (*i.e.*, the current query) from all the input visual events. Each visual event is assigned a unique index based on its chronological extraction order (*i.e.*, the first extracted visual event has an index of 1). Since all visual events are extracted sequentially, we discard audio events whose matched visual event index span exceeds 5. This filtering step excludes audio events that cover an excessively long temporal range (*e.g.*, background music that plays throughout the entire video), as such coarse-grained audio events provide limited support for fine-grained audiovisual alignment. Next, we construct a candidate visual event set containing 5 visual events by selecting those adjacent to the indices of the selected synchronized visual events.

This strategy naturally incorporates neighboring visual events, which are prone to misalignment with the audio events—thereby creating challenging hard negative samples for subsequent reward assessment. Collectively, the filtered audio events paired with their corresponding constructed visual candidate sets form the final synchronization queries $\mathcal{Q}^{\star}$.

As a result, we obtain an audiovisual event unit set and an audio-to-visual synchronization query set for each video. These well-defined reference structures enable reliable event-level reward calculation and support the subsequent GRPO training pipeline.

A.2.2 Details about the Rollout Process

SFT initializes the audiovisual captioning format of the model and enables preliminary event alignment capabilities. For GRPO, we first conduct the rollout process to identify qualified videos and yield high-quality event-level supervision signals. We perform rollout on the videos selected in Sec. A.2.1, all equipped with the well-constructed reference structures (*i.e.*, sets of audiovisual event units and audio-to-visual synchronization queries). Each selected video is then inferred 8 times (*i.e.*, group size) using the SFT model.

Leveraging the constructed reference structures, we compute two types of event-level rewards for each generated caption during the rollout process: the intra-event attribution reward and the inter-event synchronization reward. Specifically, the intra-event attribution reward focuses on evaluating the proportion of events that are correctly attributed to the appropriate characters, while the inter-event synchronization reward quantifies the alignment between audio events and the constructed candidate visual events by calculating the F1-score.

To ensure the quality of supervision signals for subsequent preference training, we implement a two-stage filtering mechanism at the query level and the video level. At the query level, invalid audio-to-visual synchronization queries are filtered by the inter-event synchronization reward. Specifically, we aggregate reward values for each individual query across the 8 generated captions, and retain only queries whose average synchronization reward falls within the interval $(0, 0.75)$, following the filtering protocol of HumanOmniV2 [55]. At the video level, filtering is performed based on the sum of intra-event attribution reward and inter-event synchronization reward. We first eliminate videos with fewer than 5 valid synchronization queries, as such cases lack sufficient synchronized semantic information to support reliable training. We then compute the standard deviation of the combined reward scores across the 8 generated captions per video, and finally retain the top 50% of qualified videos, yielding 3.2K valid video samples for downstream GRPO training.

We conduct two rounds of GRPO training, with one training epoch set for each round. In the first round, we only leverage Global Event Completeness and Inter-Event Synchronization for model optimization. Accordingly, the rollout video selection relies solely on the standard deviation of the inter-event synchronization reward. In the second round, we incorporate all three designed reward components into the joint training objective. Correspondingly, the rollout selection strategy is updated to adopt the standard deviation of the sum of inter-event synchronization reward and intra-event attribution reward as the screening criterion.

A.3 Additional Experiment Results

To ensure testing efficiency, we use Qwen3.5-27B [36] instead of Gemini-2.5-Pro [8] and Seed-1.6 [34] as the judge LLM for event-centric evaluation in the experiments in this section. Additionally, we halve the GRPO training process, retaining only 1.6K video samples after rollout.

A.3.1 Ablation on the Independent Effect of Data and Reward

In Sec. 4.3, we conduct thorough ablation experiments on our post-training pipeline. To clearly disentangle gains from constructed data versus reward guidance, we decouple the data and reward suites to evaluate each contribution to performance.

Data Effect. To separate the contribution of EVA-Data from the GRPO process, we compare SFT on EVA-Data with SFT on a same-size subset (63K) sampled from ASID-1M [21], which lacks event-centric quality control.

As shown in Tab. 5, replacing the SFT data with event-centric refined data improves both event-oriented and general performance, e.g., by +4.9% on A2V grounding and +3.6% on Daily-Omni. This demonstrates the effectiveness of our data construction pipeline and the quality of EVA-Data.

Table 5: Ablation experiments on different SFT dataset.

Dataset	V2A			A2V			AV-Speaker	Daily-Omni
	M	Acc	Score	M	Acc	Score	Acc	Acc
ASID-1M (63K)	0.9	15.2	0.8	0.9	30.7	1.8	34.1	46.5
EVA-Data	1.2	20.5	1.2	1.1	35.6	2.1	39.6	50.1

Reward Effect. To test whether our event-centric reward framework generalizes beyond EVA-Data, we apply the same training recipe to a different publicly available model, ASID-Captioner [21]. Specifically, we use its released SFT model as the baseline and then perform two stages of GRPO training on the publicly available ASID-1M dataset, using the same reward suite we proposed.

Table 6: Ablation experiments on event-centric reward framework with public SFT model and data.

Setting	V2A			A2V			AV-Speaker	Daily-Omni
	M	Acc	Score	M	Acc	Score	Acc	Acc
ASID baseline	1.1	16.8	1.0	0.9	30.4	1.8	38.4	48.5
ASID GRPO stage1	1.3	21.0	1.2	1.1	36.4	2.1	41.0	52.2
ASID GRPO stage2	1.3	22.2	1.3	1.1	39.0	2.1	42.2	53.3

These results in Tab. 6 show that, independent of EVA-Data, our event-centric framework can be effectively applied to another publicly available SFT model and yields consistent improvements across both event-centric and general benchmarks. This supports the generalizability of the reward framework and, together with the ablation in Tab. 5, further demonstrates that the final gains come from both the event-centric data and the framework.

A.3.2 Robustness of the Reward Judge

We assess the robustness of the reward judge in two ways. First, before selecting Qwen3.5-27B as the reward LLM, we sampled 200 reward-evaluation items covering the three reward dimensions and compared its judgments against Gemini-2.5-Pro. The agreement rate is around 85%, suggesting that Qwen3.5-27B provides reliable event-level judgments for reward supervision without obvious systematic bias.

Second, we also replace Qwen3.5-27B with the open-source Gemma-4-31B-it [35] as the reward judge to further test training stability under different judges. We perform GRPO training (stage 1) using the same reward design (Global Event Completeness and Inter-Event Synchronization).

Table 7: Ablation experiments on different reward LLMs.

Setting	Reward LLM	V2A			A2V			Daily-Omni
		M	Acc	Score	M	Acc	Score	Acc
SFT	-	1.2	20.5	1.2	1.1	35.6	2.1	50.1
GRPO	Qwen3.5-27B	1.4	23.0	1.3	1.2	40.1	2.3	53.0
GRPO	Gemma-4-31B-it	1.4	24.2	1.4	1.2	39.9	2.3	52.2

As shown in Tab. 7, under supervision from the Global Event Completeness and Inter-Event Synchronization rewards, both reward judges lead to similar and consistent improvements over SFT across event-centric and general benchmarks. This suggests that our reward design and training framework are robust to the choice of reward LLM, and that EVA-Cap learns fundamental event-centric alignment rather than merely exploiting the preferences of a specific judge.

A.3.3 Length-Controlled Hallucination Evaluation

All models in our evaluation use the same maximum generation length of 2048. For length-controlled evaluation, we split video-SALMONN2 testset [32] predictions by token length into [0, 700), [700, 1000), and [1000, 2048], and recompute the hallucination rate within each partition.

Table 8: Hallucination comparison under different output length.

Model	[0, 700)	[700, 1000)	[1000, 2048)
DiaDem [5]	18.6	17.7	14.9
EVA-Cap	17.4	16.8	17.0

Under this length-matched comparison, EVA-Cap shows lower hallucination than DiaDem in the short-length and mid-length splits, while DiaDem is lower in the longest split. This suggests that EVA-Cap is more reliable when producing compact, event-centric audiovisual captions, but still lags behind strong baselines on long-form outputs. We will further study the trade-off between detail and hallucination in future work.

A.3.4 Experiments with Event Precision Reward

The Global Event Completeness reward $\mathcal{R}_{\text{comp}}$ encourages more detailed captions. After GRPO training with $\mathcal{R}_{\text{comp}}$, the average caption length on video-SALMONN2 testset increases from 633.0 tokens (SFT) to 748.2 tokens. However, the per-token hallucination rate does not increase (8.2 $\rightarrow$ 8.1 Hall./1k tokens), while event coverage is substantially improved. This suggests that the higher absolute hallucination rate is largely coupled with longer, more comprehensive descriptions, rather than a pure collapse into unfaithful verbosity.

Furthermore, we add an event-level precision term to the recall-based GRPO baseline, which is computed by decomposing the predicted caption into a sequence of audiovisual events and then determining the proportion of these events that are entailed in the groundtruth caption.

Table 9: Performance comparison with event precision reward.

Reward	Avg tokens	Hall.↓	Hall./1k token↓	Miss.↓	Total↓	Daily-Omni↑
Recall	748.2	17.7	8.1	19.0	36.7	53.7
Recall + Precision	624.5	16.2	8.9	20.9	37.1	51.6

Adding precision shortens captions and reduces absolute hallucination, but increases the miss rate, slightly hurts general performance, and does not improve length-normalized faithfulness (Hall./1k: 8.1 $\rightarrow$ 8.9), while raising training cost due to extra judge calls. This indicates that precision can be viewed as encouraging more conservative outputs. In contrast, our Intra-Event Attribution reward $\mathcal{R}_{\text{attr}}$ already penalizes hallucinated or incorrectly grounded event content, providing a more targeted improvement in faithfulness. We therefore do not include precision/F1 in our reward suite.

A.4 Discussion with Recent Work

Recent studies [27, 6, 48, 28] shift caption supervision from a holistic level to fine-grained instance and event levels. Here, we focus on three concurrent works highly relevant to our own: OmniCap-Bench [56], TCA-Captioner [65], and AVSCap [47]. Targeting fine-grained audiovisual captioning, OmniCapBench formalizes video content into a structured representation and employs three interdependent tracks (*i.e.*, references, events, and shots) to evaluate whether a predicted event occurs at the correct time and links to the appropriate visual entities. TCA-Captioner emphasizes temporal and cross-modal alignment, binding auditory events to visual entities and capturing complex causal dynamics through data construction and SFT training. AVSCap achieves explicit cross-modal event binding through a designed reward, focusing on whether the caption captures temporally synchronized audiovisual events.

While our EVA-Cap shares the same objectives as these studies, we emphasize comprehensive, fine-grained, event-centric caption optimization and offer a distinct solution. Although the data curation



off-screen: … the first man in space. **Just weeks later, the United States launched its first astronaut, Alan Shepard**

Predicted Caption: The audio begins with a deep, resonant, and slightly mysterious electronic score, … cosmonaut Yuri Gagarin became the first man in space." As he speaks, the view transitions to a blurry, … and a black-and-white photograph of a man in a military uniform, Yuri Gagarin. **The narrator continues, "Just weeks later, the United States launched its first astronaut, Alan Shepard," while a graphic appears showing a line drawing of a "MERCURY-REDSTONE 3" rocket with the date "MAY 5, 1961" next to a photograph of the rocket on its launchpad.** This is followed by a black-and-white archival clip of a man in a spacesuit …



the man on the left: And today, we're going to be answering some questions from this … cookie jar?

Predicted Caption: A medium shot shows three young people … zooms in slightly as Noah, the man on the left, looks toward Ashby and continues with a friendly, engaging tone, "And today, we're going to be answering some questions from this..." As he speaks, the camera cuts to a close-up of the cookie jar, which has the word "cookies" written in green on its side, while **Ashby, the man on the right, reaches for it and finishes his sentence with a playful, slightly dramatic flourish, "...cookie jar?"** The shot returns to …

Figure 6: Failure cases of EVA-Cap, *e.g.*, videos involving rapid scene transitions and multi-character scenes with off-screen speech. Incorrect descriptions are marked in red.

pipelines differ in detail, they all achieve high-fidelity caption annotation, providing diverse options for data scaling. Methodologically, the reward design in AVSCap is closely related to our approach, paralleling the mechanism employed in our Stage 3 annotation pipeline. A key distinction in reward design is that we decouple audio and visual events for calculation during the GRPO process, whereas AVSCap treats synchronized audiovisual events as a unified whole. We leave the experimental comparison of these design choices, as well as the exploration of data mixing and methodological integration, for future work.

A.5 Failure Cases and Limitations

As shown in Fig. 6, we summarize two common failure modes of EVA-Cap: 1) In videos with rapid scene transitions, omissions in visual descriptions and misalignment between audio and visual descriptions still tend to occur. 2) In multi-character scenes with off-screen speech, the speaker identity is also prone to being incorrectly bound. This issue likely arises from limitations in understanding instances and capturing key information in complex, dynamic scenes. We intend to mitigate this limitation in future work by introducing grounded audiovisual captioning, to further enhance the fidelity and trustworthiness of audiovisual captioning by binding audio and visual cues to specific instances in space and time.

Although EVA-Data is constructed via a multi-stage pipeline (initialization–revision–filtering) that enables scalable annotation, it still relies on proprietary MLLMs for caption generation and quality control. This dependence may introduce biases from proprietary teachers and leave residual annotation errors after automatic verification. In the future, we plan to improve this pipeline by integrating agent-based, multi-level cross-verification: multiple models and rule-based checkers cross-validate captions and events at fine granularity, recursively refine inconsistent annotations, and reject insufficiently verified samples, thereby improving annotation fidelity while reducing dependence on single models.

A.6 Prompt Design

A.6.1 Prompt for Annotation Pipeline

In Figs. 7, 8, and 9, we present the prompts adopted in the audiovisual captioning data annotation pipeline introduced in Sec. 3.1.1. In this process, we first prompt Gemini-2.5-Pro to refine the initial audiovisual captions in terms of cross-modal synchronization and event attribution, and to extract timestamped audiovisual event tuples in parallel. We then employ GPT-4.1 to conduct event hit verification for data quality control.

SYSTEM ROLE
You are a professional Multimodal Video Audit and Text Refinement Expert. You specialize in synchronizing complex audiovisual signals and ensuring precise identity-event binding in video narratives.

INPUT DATA
- Video File: An MP4 file containing both visual and audio streams.
- Original Detailed Caption: A preliminary video description generated by an Omni-model.

TASK
Refine and correct the provided Detailed Caption based on the actual video and audio content. Your goal is to produce a high-fidelity narrative where every sound and action is perfectly aligned and attributed.

CORE ALIGNMENT REQUIREMENTS (PRIORITY)
1. **Audiovisual Sync**: If a visual description and an audio (ASR/sound) event occur together, you MUST use explicit markers (e.g., "simultaneously," "while," "as") to bind them. Do not describe concurrent events as a sequence.
2. **Event Attribution**: Every action, speech, and sound MUST be attributed to a specific, identifiable subject using unique visual traits (e.g., "the man in the blue shirt"). For off-screen voices, identify them as a "narrator" with a brief vocal description.

GENERAL REFINEMENT RULES
3. **Redundancy Elimination**: If the ASR content is identical to the on-screen captions (OCR) at any given moment, **KEEP ONLY the spoken dialogue** and do not describe the on-screen text separately.
4. **Accuracy Correction**: Directly correct any OCR or ASR errors (misspellings, wrong words, or hallucinations) present in the Original Caption.
5. **Style Preservation**: Maintain the narrative style and tone of the original description. **Do NOT discard** any accurate visual or environmental details.

OUTPUT CONSTRAINTS (STRICT)
- **Single Paragraph**: Present the refined caption as one continuous, cohesive paragraph.
- **No Meta-talk**: Return ONLY the corrected text. No introductory phrases like "Here is the refined version," "Updated caption," or conversational filler.

START REFINEMENT
Here is the Original Detailed Caption: """*A draft video caption describing the visible people, actions, setting, spoken words, and background sounds that may need correction or synchronization.*"""
Now generate the Refined Detailed Caption:

Figure 7: Prompt for caption refinement during annotation stage 2 with Gemini-2.5-Pro.

SYSTEM ROLE
You are a professional Multimodal Video Event Analyst. Your task is to extract a sequence of synchronized event segments from a video, focusing only on significant and meaningful occurrences.

INPUT DATA
- Video File: An MP4 file containing visual and audio streams.

TASK
Analyze the video and output a list of Synchronized Event Triples (Time Range, Audio Event, Visual Event). **Note**: Time segments MAY overlap. If a continuous event in one modality spans multiple distinct events in the other, repeat the description as needed to ensure each segment is a complete, synchronized unit.

JSON SCHEMA
Output the data strictly in the following format:

```
[
  {
    "start_time": "SS.ms",
    "end_time": "SS.ms",
    "audio_event": "A single complete sentence describing the subject and the sound/ASR, or 'None'.",
    "visual_event": "A single complete sentence describing the subject and the action/background/OCR."
  }
]
```

EXTRACTION REQUIREMENTS
1. **Saliency-Based Segmentation**: Extract only significant segments where a meaningful event begins and ends, rather than partitioning the entire video.
2. **Audio-Visual Alignment**: Ensure that the audio_event and visual_event are strictly concurrent and physically synchronized within the same start_time and end_time.
3. **Audio Narrative**: Describe the speech or sound in one complete sentence. You MUST identify the specific subject (e.g., "narrator" with a brief vocal description for off-screen voice, or a descriptive trait for on-screen subjects) followed by their ASR content or sound type. For ASR, provide a full transcribed sentence and ignore brief natural pauses within the sentence to maintain semantic integrity. Use "None" if there is no significant audio.
4. **Visual Narrative**: Describe the smallest semantic unit of a physical action or state change in one complete sentence. Do not combine multiple distinct actions; each segment must focus on a single movement or environmental shift (e.g., "The text [content] is visible on the right" or "The camera zooms into the building"). You MUST identify the specific subject followed by the action. Strictly ignore ASR-matching subtitles.

OUTPUT CONSTRAINT
Return ONLY the raw JSON list. No introductory text or meta-commentary.

Figure 8: Prompt for timestamped event tuple extraction during annotation stage 3 with Gemini-2.5-Pro.

SYSTEM ROLE
You are a professional Multimodal Logical Inference Expert. Your task is to determine whether a series of "Synchronized Event Triples" are fully supported and entailed by a provided **Video Omni Caption**.

INPUT DATA
- **Video Omni Caption**: A comprehensive description of the video content.
- **Synchronized Event Triples**: A list of objects: `[{"start_time": "SS.ms", "end_time": "SS.ms", "audio_event": "...", "visual_event": "..."}, ...]`

TASK STEPS
1. **Independent Verification**: For each event triple, separately verify the audio and visual components against the Video Omni Caption:
   - **Audio Hit**: Does the caption entail the specific **subject and sound/ASR**? If `audio_event` is "None", mark "Yes" (as the absence of a specific event is the target to verify in the next step).
   - **Visual Hit**: Does the caption entail the specific **subject and action/background/OCR**?
2. **Synchronization Inference**: Determine if the caption confirms these two events occurred **simultaneously**:
   - **Standard Case**: If `audio_event` is NOT "None", mark `is_synchronized` as "Yes" ONLY if the caption explicitly or logically shows both the audio and visual events happening at the same time (e.g., linked by "while", "as", "during", or in the same sentence).
   - **None-Audio Case**: If `audio_event` is "None", mark `is_synchronized` as "Yes" ONLY if the visual event is **solitary**. If the caption mentions any speech or significant sound occurring during this visual action, mark "No".
3. **One-to-One Mapping**: Process every item in the input list and maintain the original order.

OUTPUT FORMAT (STRICT JSON LIST)
Return the results strictly as a JSON **LIST** of objects. Do not include any introductory phrases or conversational filler.
Format:

```
[
  {
    "start_time": "Original start_time",
    "end_time": "Original end_time",
    "audio_event": "Original audio_event",
    "visual_event": "Original visual_event",
    "audio_hit": "Yes/No",
    "visual_hit": "Yes/No",
    "is_synchronized": "Yes/No",
    "reason": "Provide a concise justification: 1. Cite the caption evidence for audio and visual hits. 2. Explain how temporal synchronization is proven or why it fails."
  }
]
```

START ANALYSIS
Here is the Video Omni Caption: """*A comprehensive Omni caption for the full video, including the chronological visual actions, identifiable subjects, scene setting, camera movement, non-redundant OCR, spoken dialogue or ASR, sound effects, background audio, and any wording that indicates whether audio and visual events occur together or separately.*"""
Here are the Synchronized Event Triples: *A JSON list of candidate synchronized event triples, where each item contains a start time, end time, audio event, and visual event that must be checked against the caption for subject-level support, visual evidence, audio evidence, and temporal alignment.*

Figure 9: Prompt for event-centric data verification during annotation stage 3 with GPT-4.1.

A.6.2 Prompt for Reward Implementation

Fig. 10 illustrates the prompt used to extract audiovisual events from audiovisual captions, facilitating the construction of reference structures. Furthermore, Figs. 11, 12 and 13 list the prompts used, respectively, for the judgement of global event completeness reward, intra-event attribution reward and inter-event synchronization reward. All these prompts are carefully designed to standardize LLM-based reward judgement and ensure consistent evaluation across all samples.

A.7 Broader Impact

EVA-Cap aims to advance audiovisual video captioning by generating more faithful, comprehensive, and temporally coherent descriptions of videos. This can benefit various downstream applications that rely on detailed and accurate video understanding. Better alignment between audio and visual events may help reduce misleading descriptions in scenarios where both modalities are necessary for understanding the video content.

At the same time, improved audiovisual captioning systems may introduce potential risks if deployed without appropriate safeguards. First, generated captions may still contain omissions, hallucinations, incorrect speaker attributions, or inaccurate audio-visual temporal relations. Such errors could mislead users if the captions are treated as factual summaries. Second, automatic captioning of user-generated videos may raise privacy concerns when videos contain identifiable people, sensitive conversations, or private environments. Third, high-quality video descriptions could be misused for large-scale surveillance, profiling, or automated indexing of sensitive content.

You are a professional video/audio analysis expert. Your task is to extract significant, time-ordered dynamic events from the provided [Caption] into a structured JSON format.

Rules:
1. **Chronological Order**: Events must be listed in the exact order they occur in the caption.
2. **Audio Logic**:
   - Types: "asr" (spoken words), "sound" (physical sounds, ambient effects, etc.).
   - Character: The speaker's description (e.g., "Off-screen narrator"). Use "N/A" for general sounds.
   - Content (ASR): Verbatim transcript. STRICTLY EXCLUDE filler words (e.g., "uh", "um", "oh").
   - Content (Sound): A complete sentence describing the sound (e.g., "A dog barks loudly").
3. **Visual Logic (Max 20 entries)**:
   - Types: "action" (physical movement), "camera" (pan/zoom/cut/angle), "environment" (text appearance, background).
   - Character: The specific discriminative agent (a person or an animal) performing the action. Use "N/A" for camera or environment types.
   - Content: A single, complete standalone sentence describing the change (e.g., "The cat jumps onto the table").
4. Constraint: Do not treat "speaking" as a visual action; it belongs in the audio section.
5. **Output Format (JSON)**: Return ONLY a valid JSON object with the following structure. Do not include any conversational text or Markdown code blocks.

```
{
  "audio": [{"modal": "audio", "type": "asr/sound", "character": "...", "content": "..."}, ...],
  "visual": [{"modal": "visual", "type": "action/camera/environment", "character": "...", "content": "..."}, ...]
}
```

Caption:
A detailed Omni caption containing the chronological visual actions, scene changes, OCR when relevant, spoken words, sound effects, ambient audio, and subject descriptions from the video.

Output (JSON):

Figure 10: Prompt for audiovisual event extraction with Qwen3.5.

You are a professional factual verification expert. Your task is to verify whether the [Target Event] is logically entailed and supported by the [Caption].

Rules:
1. **Modality-Specific Verification**:
   - **If modal is "audio"**: Focus on what is SAID or HEARD. The exact spoken words (ASR) or specific sound descriptions must be present in the [Caption].
     - **Note**: If type is "asr", use fuzzy matching. Allow homophones (same sound, different spelling), minor ASR errors, or slight variations in phrasing (e.g., "I'm going" vs "I am going"). The core meaning and key nouns must match.
   - **If modal is "visual"**: Focus EXCLUSIVELY on physical ACTIONS, movements, and visual appearances.
     - **Note**: If the [Target Event]'s content for a visual modal includes descriptions of speaking, tone of voice, or dialogue (e.g., "says while jumping", "jumps angrily"), **IGNORE the speaking/tone part** and only verify the physical action (e.g., "jumping").
2. **Fact-Checking Logic**:
   - **"Yes"**: The [Target Event]'s content is explicitly stated by the [Caption]. Synonym matches are allowed (e.g., "running" vs "jogging").
   - **"No"**: The [Caption] lacks this specific information, or provides contradictory information.
3. **Strict Constraints**:
   - Base your judgment *only* on the [Caption].
   - If the [Target Event] contains multiple details, all details must be supported for a "Yes".
4. **Output Format**:
   - Return ONLY a JSON object: {"modal": "...", "type": "...", "content": "...", "verified": "Yes/No"}.
   - After the JSON, provide a brief reasoning paragraph explaining the decision.
   - Do not include any conversational filler.

Caption:
A detailed Omni caption containing the available evidence about visible actions, appearances, scene context, OCR, spoken words, sound events, and any relevant temporal or causal relationships.

Target Event:
A single event dictionary with modal, type, character or subject information when available, and content that must be verified against the caption.

Output (JSON):

Figure 11: Prompt for global event completeness reward judgement via Qwen3.5.

You are a factual consistency auditor. Your goal is to verify if the [Target Event] is accurately attributed to the correct [Character] in the [Caption] through a multi-step logical deduction.

Rules:
1. **Step 1: Event Localization (Content Match)**
  - Scan the [Caption] to locate the specific event or speech that matches the [Content].
  - **Guidance**: This event is confirmed to exist. Identify the specific segment where it occurs.
  - **Modality-Specific**:
    - For "audio": Focus on what is SAID or HEARD. Use fuzzy matching for "asr" type.
    - For "visual": Focus EXCLUSIVELY on physical ACTIONS, movements, and visual appearances. Ignore any mentioned tone or descriptions of speaking.
2. **Step 2: Global Character Profiling (Character Extraction)**
  - **Trace**: Identify the immediate subject (performer) of the event found in Step 1.
  - **Aggregate**: Scan the entire [Caption] to collect all descriptors (role, attire, gender) linked to this character via anaphora (e.g., "he", "she", "the man").
  - **Note**: For "asr" type, if the character is an "Off-screen narrator," focus on vocal descriptors to build the profile.
  - **Strict Rule**: If the [Caption] provides no identifying features for this character, report it as "Generic/Missing Profile". Do NOT hallucinate.
3. **Step 3: Entailment Judgment**
  - Compare the "Character Profile" (from Step 2) with the [Character] phrase from the Target Event.
    - **Yes**: The profile is logically consistent with and supports the [Character] description.
    - **No**: There is a conflict (wrong gender, wrong clothing) or the description is so vague (e.g., "someone", "a person") that it fails to uniquely identify the [Character].
4. **Output Format**:
  - Return ONLY a JSON object: {"modal": "...", "type": "...", "character": "...", "content": "...", "verified": "Yes/No"}.
  - After the JSON, provide a brief reasoning paragraph explaining the decision:
    - **Character Profile**: List descriptions of this character found in caption.
    - **Entailment Logic**: Explain why the profile matches or fails to match the target character.

Caption:
A detailed Omni caption used as the only evidence source, including the event context, subject references, pronouns, descriptors, clothing, roles, gender cues, vocal descriptors, and any anaphora needed to profile the character.

Target Event:
A single event dictionary containing the modal, type, target character phrase, and event content whose attribution must be checked against the character profile inferred from the caption.

Output (JSON):

Figure 12: Prompt for intra-event attribution reward judgement via Qwen3.5.

You are a precise Audio-Visual Synchronization Auditor. Your task is to determine which [Visual Candidates] occur at the same time as the [Audio Event], using the [Caption] as the sole source of truth.

Rules:
1. **Step 1: Audio Identification (The Gatekeeper)**:
  - Scan the [Caption] for the [Audio Event].
  - **If Type is "asr"**: Use fuzzy matching. Allow homophones (same sound, different spelling), minor ASR errors, or slight variations in phrasing (e.g., "I'm going" vs "I am going"). The core meaning and key nouns must match.
  - **If Type is "sound"**: Use semantic matching. Focus on the nature of the sound (e.g., "A dog barks" matches "Canine barking" or "The puppy yelps").
2. **Step 2: Individual Candidate Audit**
  For each candidate in the [Visual Candidates] list, it is ONLY considered synchronous if it meets BOTH conditions:
  - **Condition A (Temporal Connection)**: The visual event and the audio event are explicitly linked by words like "as", "while", "simultaneously", or described as a concurrent gesture (e.g., "says ..., gesturing with hands").
  - **Condition B (No Cut Barrier)**: There must be NO scene-change or camera-cut markers (e.g., "cuts to", "the scene shifts to", "returns to") between the audio description and the visual event in the [Caption]. A cut indicates a new timestamp.
3. **Step 3: Verification**
  - Return all indices that pass both conditions.
  - If the [Audio Event] cannot be found, or no visual candidate passes both conditions, return "N/A".
4. **Output Format**:
  - Return ONLY a JSON object: {"selected_indices": a list of integers (e.g., [1, 3]) or "N/A"}.
  - After the JSON, provide a brief reasoning paragraph explaining the decision:
   - **Audio Identification**: Compare [Audio Event] quote vs. [Caption] quote.
   - **Sync Logic**: Explain why the indices were selected or why "N/A" was chosen.

Caption:
A detailed Omni caption used as the only evidence source, including the ordering of audio descriptions, visual events, scene cuts, camera shifts, concurrent wording, and any explicit synchronization markers.

Audio Event:
The target audio event to locate in the caption, including its type such as ASR or sound and its spoken content or sound description.

Visual Candidates (Numbered 1 to N):
A numbered list of candidate visual events whose temporal relationship to the audio event must be audited one by one.

Output (JSON):

Figure 13: Prompt for inter-event synchronization reward judgement via Qwen3.5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly and accurately state the paper's contributions and scope in the abstract and introduction (Section 1).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the EVA-Data construction and quality-control pipeline in Section 3.1, and the main implementation details, hyperparameters, and evaluation settings in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code, model checkpoints, and reproducible data construction scripts upon publication. For data that cannot be redistributed directly, we will provide processing instructions and metadata to facilitate reproduction.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We specify implementation details, hyperparameters, and evaluation settings in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars due to the high computational cost of repeated large-scale omni-modal training runs. We instead support the main conclusions through consistent improvements across multiple benchmarks and ablation settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies details about computer resources in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and conducted the research in accordance with its guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Appendix A.7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not implement additional release safeguards beyond standard documentation and usage guidance. We discuss potential misuse and privacy risks in the broader impact statement and will follow the licenses and privacy requirements of the underlying data sources when releasing assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the third-party datasets, pretrained models, and benchmarks used in this work, and will respect their licenses and terms of use when releasing code, models, or derived data.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This work introduces EVA-Data and trained EVA-Cap checkpoints as new assets. We describe the data construction pipeline, quality-control procedure, data statistics, training settings, and evaluation protocols in the paper, and will provide documentation, licenses, and usage instructions with the released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The use of LLMs and omni-modal language models is central to our data construction, reward design, and training framework, as described in Section 3.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.