PIGEONS: INCREMENTAL LONG-VIDEO CAPTIONING WITH PERSISTENT IDENTITY GALLERY AND EVOLVING OVERARCHING NARRATIVE SUMMARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Detailed video captioning has become an essential semantic interface for multimodal understanding and generation. While existing methods mainly target short videos, extending them to long videos remains challenging, primarily due to limited context windows of current video understanding models. A naive solution is to adopt segment-wise captioning followed by concatenation, yet this paradigm often suffers from inconsistent subject references and fragmented long-range narratives. To mitigate these issues, we introduce **PIGEONS**, an incremental long-video captioning model built upon a Persistent Identity Gallery and an Evolving Overarching Narrative Summary. Before caption generation, PIGEONS scans the entire video to construct an identity gallery that captures key subjects along with their visual references, thereby providing a stable referential anchor. The model then performs clip-by-clip captioning conditioned on the persistent gallery and an evolving overarching narrative summary, which compactly maintains and updates essential long-range contextual information after each captioning step. To enable these capabilities, we develop a three-stage training strategy that integrates supervised fine-tuning and reinforcement learning. Extensive experiments demonstrate that PIGEONS significantly improves long-video captioning performance, achieving superior accuracy, completeness, and subject referential consistency over existing approaches, while maintaining competitive performance on short-video captioning. PIGEONS will be open-sourced to facilitate future research.

1 INTRODUCTION

In the era of multimodal large language models (MLLMs), detailed video captioning has emerged as a crucial semantic interface that bridges complex perceptual signals with fine-grained linguistic representations (Chai et al., 2025; Chen et al., 2024; 2026a). High-quality detailed video captions not only facilitate effective alignment among audio, visual, and textual modalities during pretraining (Xu et al., 2025b; Team et al., 2025), but also provide rich semantic knowledge for downstream tasks (Long et al., 2026; Du et al., 2025). Recent studies have consistently demonstrated that enhancing the quality of detailed video captions leads to notable performance gains across diverse multimodal applications (An et al., 2026; Team, 2026; Zhou et al., 2026; Liu et al., 2026).

Despite these advancements, existing research primarily focuses on relatively short videos (Wang et al., 2024; Yuan et al., 2025; Xu et al., 2026; Zhang et al., 2026), leaving long-form videos, which more closely resemble real-world scenarios, largely underexplored. A key challenge lies in the limited context windows of current video understanding models, which make it intractable to process an entire long video in a single pass while preserving sufficient temporal and spatial resolution for fine-grained understanding. To address this, a straightforward approach is segment-wise captioning, where a long video is divided into shorter clips, each clip is captioned independently, and the resulting clip-level captions are concatenated to form a complete video-level description.

However, this naive segment-wise approach introduces two critical limitations: (1) *loss of subject referential consistency*, and (2) *fragmented long-range narratives*. First, the same subject may be referred to using varying descriptions across clips. While locally accurate, the lack of stable identity associations makes it difficult to resolve coreferences in the final concatenated caption, thus con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

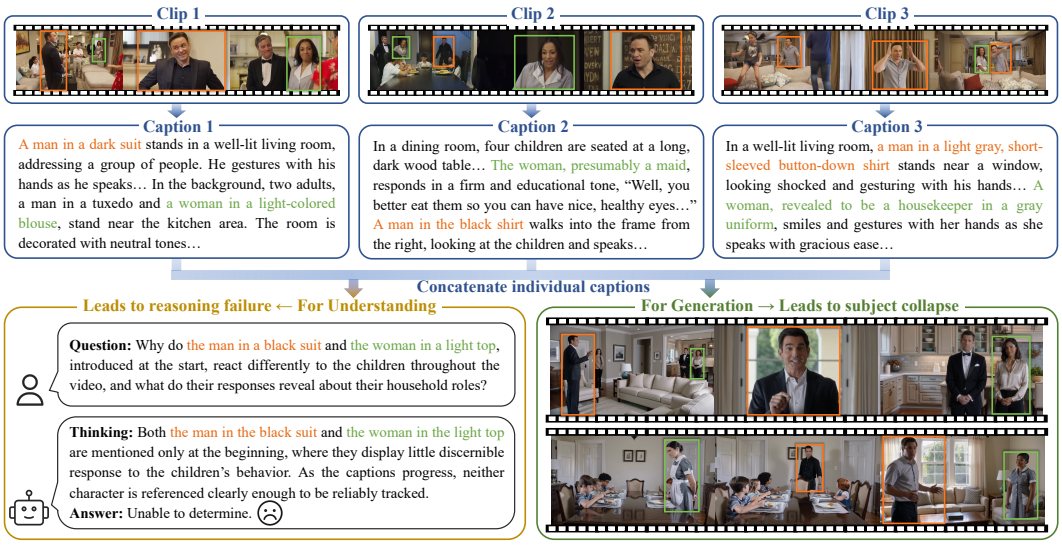


Figure 1: Downstream impacts of inconsistent subject references in naive segment-wise captioning. In understanding tasks, using such captions as the memory of an LLM agent may induce reasoning failures due to identity ambiguity. In generation tasks, leveraging these captions for video reconstruction could lead to subject collapse. Identical colors represent the same ground-truth subject.

fusing character relationships, disrupting narrative flow, and ultimately degrading the performance of downstream caption-based understanding and generation tasks, as shown in Figure 1. Second, isolated clip captioning ignores historical context. Models may redundantly reintroduce previously observed subjects or scenes, overlook established facts, or treat causally connected events as isolated local actions. While a straightforward remedy is to append all preceding captions as context, this approach scales poorly. As the video progresses, the accumulated caption history grows rapidly, which not only incurs substantial computational overhead but also introduces irrelevant details that distract the model from accurately describing the current event.

To mitigate these limitations, we introduce **PIGEONS**, an incremental long-video captioning model designed around a Persistent Identity Gallery and an Evolving Overarching Narrative Summary. Specifically, prior to segment-wise captioning, PIGEONS scans the entire video to construct a global identity gallery, assigning unique identifiers and reference images to key subjects. This gallery remains fixed throughout the subsequent clip-by-clip captioning process, serving as a stable anchor for global subject referential consistency. After captioning each clip, PIGEONS simultaneously extracts essential elements for subsequent narrative understanding (e.g., existing scenes, character dynamics, subject roles, and established facts) and integrates them into an evolving overarching narrative summary. While this summary captures long-range narrative context compactly, the caption of the immediately preceding clip is retained to ensure local continuity. Consequently, when generating the caption for the n-th clip ($n > 1$), PIGEONS leverages three complementary sources of context: the persistent identity gallery for subject referential consistency, the evolving overarching summary for long-range narrative coherence, and the caption of the $(n-1)$-th clip for local continuity.

Equipping the model with these capabilities poses a dual challenge: identity-gallery construction requires precise spatiotemporal localization to identify representative subject images, while long-video captioning demands a comprehensive understanding of both video content and multimodal context. To tackle this, we design a tailored three-stage training strategy. Stage 1 applies supervised fine-tuning (SFT) on 16K identity-gallery construction samples, enabling preliminary capabilities in subject discovery and reference-image extraction. Stage 2 employs reinforcement learning (RL) with specialized reward functions to boost both gallery accuracy and construction efficiency. Stage 3 conducts SFT on 16K long-video captioning samples, training the model to generate multi-context-conditioned captions and dynamically update the evolving overarching summary. To mitigate catastrophic forgetting, identity-gallery construction data is mixed into this final stage. By jointly optimizing identity-gallery construction and context-aware captioning, PIGEONS is able to produce long-range cohesive captions with robust subject referential consistency.

Experimental results reveal that PIGEONS significantly outperforms existing models in long-video captioning, achieving superior accuracy, completeness, and subject referential consistency. Compared with naive segment-wise captioning, PIGEONS produces more coherent captions with stronger referential consistency, higher information density, and smoother cross-clip transitions. In addition, PIGEONS achieves highly competitive performance in short-video captioning scenarios that do not require segment-wise processing. Our contributions are summarized as follows:

- We introduce PIGEONS, an incremental long-video captioning model that leverages a persistent identity gallery and an evolving overarching narrative summary to ensure robust subject referential consistency and long-range narrative coherence.
- We develop a three-stage training strategy combining SFT and RL to jointly optimize precise identity-gallery construction and context-aware long-video captioning capabilities.
- Extensive experiments demonstrate that PIGEONS outperforms all existing models in terms of accuracy, completeness, and subject referential consistency for long-video captioning, while also delivering competitive performance in short-video captioning tasks.

2 Related Work

2.1 Detailed Video Captioning with MLLMs

Recent advances in MLLMs have propelled detailed video captioning. Early efforts focus on short, visual-only clips spanning tens of seconds, emphasizing fine-grained coverage and accuracy (Wang et al., 2024; Yuan et al., 2025; Chai et al., 2025; Zhong et al., 2026). Subsequent works incorporate auditory signals and extend target durations to the minute scale, enabling richer multimodal descriptions (Tang et al., 2025; Wu et al., 2025; Geng et al., 2025; Chen et al., 2026a; Ma et al., 2026; Li et al., 2026). Nevertheless, these time horizons still fall short of real-world long-form videos.

To handle long videos, existing research typically adopts a hierarchical paradigm (Islam et al., 2024; Wei et al., 2025; Chu et al., 2025), decomposing videos into short clips for independent captioning, followed by global summarization. However, such isolated processing ignores cross-clip context, inevitably causing inconsistent subject references and fragmented narratives. Several works have attempted to alleviate these issues. StoryTeller (He et al., 2024) uses external movie cast information to assist subject identification across clips, but such auxiliary annotations are rarely available for general videos, and the simply concatenated captions still lack strong narrative coherence. RICE (Yang et al., 2025) explores subject association by prompting models to describe subject features, enabling language-based subject matching across captions. However, such textual association lacks stable visual anchors to reliably maintain referential consistency throughout long videos.

Distinct from these methods, PIGEONS explicitly maintains both subject identity and narrative context during incremental long-video captioning. By establishing a persistent visual identity gallery as a stable reference anchor, and continuously updating an overarching summary to preserve historical context, PIGEONS ensures consistent subject references and coherent long-range storytelling.

2.2 Reinforcement Learning for Video Captioning

Reinforcement learning (Schulman et al., 2017; Shao et al., 2024; Zheng et al., 2025; Yu et al., 2026) has been widely applied to video captioning with diverse reward designs. VideoChat-R1 (Li et al., 2025) employs event-recall rewards to improve caption completeness, while VideoCap-R1 (Meng et al., 2025) decomposes captioning into structured reasoning and generation stages and leverages dual scorers to elevate output quality. CapRL++ (Yang et al., 2026) uses the performance of an LLM on video-related question answering based on generated captions as an indirect reward signal for caption optimization. OmniScript (Pu et al., 2026) optimizes long-form cinematic script generation using a temporally segmented reward that evaluates event-level accuracy across multiple dimensions.

Different from these approaches that directly optimize specific aspects of textual quality, PIGEONS targets the challenge of subject referential consistency in segmented long-video captioning. Instead of refining caption generation itself, PIGEONS leverages RL to optimize the upstream construction of the identity gallery, producing more accurate identity representations that provide explicit visual anchors to maintain subject referential consistency during subsequent captioning process.

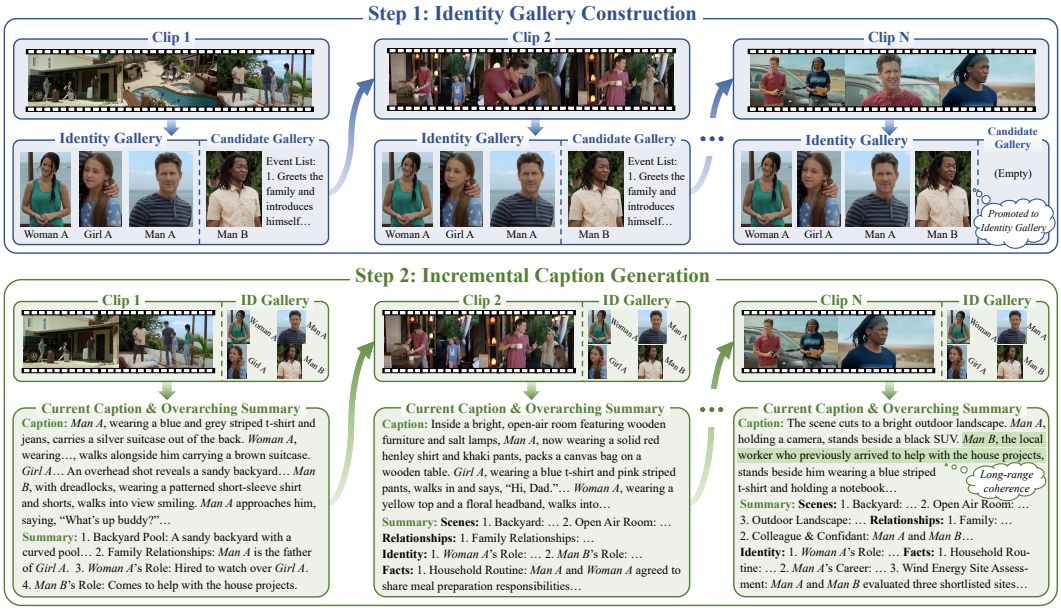


Figure 2: Inference pipeline of PIGEONS. The model first scans all video clips to construct a global identity gallery. Guided by this gallery, it sequentially generates captions for each clip, with consistent subject references in *italic*. Concurrently, an overarching narrative summary is dynamically updated to help enhance long-range coherence and information density of the generated captions.

3 PIGEONS

3.1 Overview

To mitigate inconsistent subject referencing and fragmented long-range narratives caused by naive segment-wise captioning, we introduce **PIGEONS**. As illustrated in Figure 2, PIGEONS enhances the standard segment-wise captioning paradigm by integrating two key components: a *persistent identity gallery* that provides visual subject reference anchors, and an *evolving overarching narrative summary* that offers compact global narrative context. To enable these capabilities, we further develop a tailored three-stage training strategy combining SFT and RL. Details are provided below.

3.2 Persistent Identity Gallery

To maintain subject referential consistency across video clips, PIGEONS constructs a global *identity gallery* prior to caption generation. This gallery stores unique identifiers and reference images for key subjects, and remains fixed throughout the subsequent captioning process to serve as a stable visual anchor for consistent subject referencing.

We define a *subject* as an entity (not limited to human characters) that drives the plot and plays a significant role in the video's narrative. Recognizing that an entity may appear sparingly in early clips but become crucial later, we maintain an auxiliary *candidate gallery* alongside the identity gallery. The candidate gallery tracks potential subjects in early stages along with their associated key events. Once an entity's key-event count reaches a predefined threshold (empirically set to 4), it is promoted to the identity gallery, as illustrated by *Man B* in the upper portion of Figure 2.

Formally, for a video V partitioned into N sequential clips $\{v_1, v_2, \ldots, v_N\}$, let $\mathcal{G}_I^k$ and $\mathcal{G}_C^k$ represent the identity gallery and candidate gallery obtained after processing up to the k-th clip v_k, respectively. Initially, both galleries are empty, i.e., $\mathcal{G}_I^0 = \emptyset$ and $\mathcal{G}_C^0 = \emptyset$. At the k-th step, the model $\mathcal{M}$ takes the current clip and existing galleries as inputs to yield the updated galleries:

$$(\mathcal{G}_I^k, \mathcal{G}_C^k) = \mathcal{M}(v_k, \mathcal{G}_I^{k-1}, \mathcal{G}_C^{k-1}). \tag{1}$$

After processing all N clips, the final identity gallery $\mathcal{G}_I^N$ is fixed and serves as a persistent global visual anchor to ensure consistent subject referencing throughout the subsequent captioning process.

3.3 Evolving Overarching Narrative Summary

To leverage historical context without incurring context window explosion, PIGEONS maintains an *evolving overarching narrative summary* that is updated after generating each clip's caption. Rather than naively accumulating verbose past captions, the model compresses the global narrative across four critical dimensions: *existing scenes*, *character relationships*, *subject roles*, and *established facts*. This multidimensional summarization preserves essential high-level semantic cues for understanding subsequent clips, ultimately yielding captions with enhanced long-range narrative coherence and higher information density, as depicted in the lower half of Figure 2.

While the evolving summary provides global context, the caption of the immediately preceding clip is additionally retained to maintain local narrative fluency. Formally, let c_k and s_k denote the generated caption and the updated narrative summary for clip v_k, with c_0 and s_0 initialized as empty strings. For the k-th clip, the model $\mathcal{M}$ generates its caption and updates the summary conditioned on the global identity gallery $\mathcal{G}_I^N$, the current video clip v_k, the preceding caption c_{k-1}, and the previous summary s_{k-1}:

$$(c_k, s_k) = \mathcal{M}(\mathcal{G}_I^N, v_k, c_{k-1}, s_{k-1}). \tag{2}$$

The final long-video caption c_{video} is then obtained by concatenating all clip-level captions:

$$c_{\text{video}} = \text{Concat}(c_1, c_2, \ldots, c_N). \tag{3}$$

3.4 Three-Stage Training

Building a reliable identity gallery requires precise spatiotemporal localization to identify representative subject images, while long-video captioning demands a comprehensive understanding of both video content and multimodal context. To equip our base model, Qwen3-Omni-Instruct (Xu et al., 2025b), with these dual capabilities, we design a tailored three-stage training strategy.

3.4.1 Stage 1: Identity Gallery SFT (IG-SFT)

In the first stage, we curate a dataset of 16K identity gallery construction samples and perform SFT to teach the model to identify subjects in videos and determine the appropriate timestamps and spatial bounding boxes that capture discriminative and representative views of each subject. Considering that even minor temporal or spatial deviations can substantially degrade the quality of reference images, we explicitly incorporate a verification and refinement procedure, encouraging the model to iteratively self-evaluate and correct low-quality crops, thereby producing higher-quality identity galleries, as illustrated in the left part of Figure 3.

3.4.2 Stage 2: Identity Gallery RL (IG-RL)

During identity gallery construction, the critical timestamp and bounding box information occupies only 2.7% and 3.6% of the total output sequence, making token-level SFT insufficient for providing dense supervision over these key decisions. Therefore, in the second stage, we introduce RL, specifically Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025), to further optimize gallery construction. As illustrated in the upper right part of Figure 3, we design two outcome-oriented reward functions targeting gallery quality and construction efficiency, respectively, and apply RL on a randomly sampled subset of 2K instances from the preceding SFT stage.

Identity Gallery Accuracy Reward R_{acc}. Since the construction process of an identity gallery is not uniquely determined, we define R_{acc} based on the similarity between the predicted and the ground-truth (GT) identity galleries. Formally, for a video segmented into N clips, the model ultimately constructs an identity gallery $\mathcal{G}_I^N = \{(s_p, r_p)\}_{p=1}^P$ consisting of P subject-reference pairs, and the corresponding GT identity gallery is denoted as $\tilde{\mathcal{G}}_I = \{(\tilde{s}_q, \tilde{r}_q)\}_{q=1}^Q$, containing Q pairs, where s denotes the subject identifier and r represents the corresponding reference image.

To measure the similarity between two galleries, we first define the pairwise similarity between their entries (s_p, r_p) and $(\tilde{s}_q, \tilde{r}_q)$. Since subject identifier s is merely a specific instantiation without intrinsic semantic correspondence, while the reference image r directly determines the quality of identity representation, we measure entry-level similarity based solely on the reference images. Furthermore, as an identity gallery is inherently an unordered set, we formulate the gallery-level similarity as a maximum-weight bipartite matching problem with a one-to-one assignment constraint.

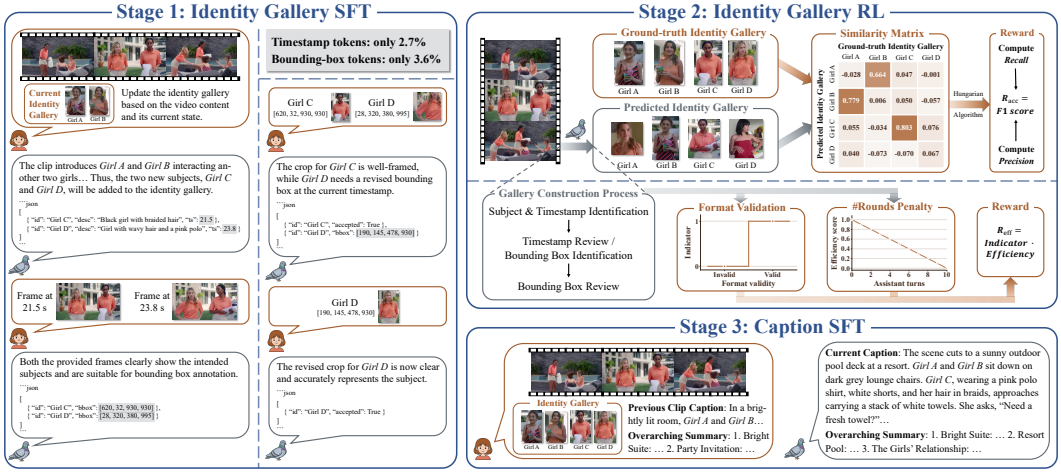


Figure 3: Overview of the three-stage training strategy. Stage 1 conducts SFT for identity gallery construction with self-review and refinement. Stage 2 further optimizes gallery construction via RL, improving both accuracy (R_{acc}) and efficiency (R_{eff}). Stage 3 performs caption SFT, enabling detailed caption generation conditioned on multiple contexts while evolving the overarching narrative summary. For clarity, only one clip from a full video is shown and the candidate gallery is omitted.

Specifically, we first employ the face detection confidence score $\text{Conf}(\cdot)$ from ArcFace (Deng et al., 2019) to determine whether a reference image contains a recognizable face. For two face-containing images, we compute their similarity using ArcFace. For two non-face images, we use DINOv3 (Siméoni et al., 2025) to measure their visual similarity. The similarity between a face-containing image and a non-face image is set to zero. Formally:

$$\text{Sim}(r_p, \tilde{r}_q) = \begin{cases} \text{ArcFace}(r_p, \tilde{r}_q), & \text{if } \text{Conf}(r_p) > \gamma \text{ and } \text{Conf}(\tilde{r}_q) > \gamma, \\ \text{DINOv3}(r_p, \tilde{r}_q), & \text{if } \text{Conf}(r_p) \le \gamma \text{ and } \text{Conf}(\tilde{r}_q) \le \gamma, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where the face confidence threshold γ is empirically set to 0.35.

Based on the pairwise similarities, we build a similarity matrix $S \in \mathbb{R}^{P \times Q}$ between the predicted and GT galleries. To avoid unreliable assignments caused by weakly correlated pairs, we apply metric-specific thresholds: S_{pq} is set to $\text{Sim}(r_p, \tilde{r}_q)$ only if the similarity exceeds $\tau_{\text{arc}} = 0.35$ for face pairs or $\tau_{\text{dino}} = 0.55$ for non-face pairs; otherwise, S_{pq} is set to zero. The Hungarian algorithm (Kuhn, 1955) is then applied to find the globally optimal bipartite matching, and the total similarity score S_{match} is given by:

$$S_{\text{match}} = \max_{\mathcal{H}} \sum_{(p,q)\in\mathcal{H}} S_{pq}, \tag{5}$$

where $\mathcal{H}$ denotes a valid bipartite matching under the one-to-one assignment constraint.

From a physical interpretation, S_{match} represents the quality of the reference images in $\mathcal{G}_I^N$, which takes values in the range $[0, \min(P, Q)]$. The precision and recall are then computed as:

$$\text{Pre} = S_{\text{match}}/P, \qquad \text{Rec} = S_{\text{match}}/Q. \tag{6}$$

Ultimately, the overall matching quality between the two identity galleries is quantified by the F_1 score, which serves as the identity gallery accuracy reward:

$$R_{\text{acc}} = 2 \cdot \text{Pre} \cdot \text{Rec}/(\text{Pre} + \text{Rec}). \tag{7}$$

Efficiency-Aware Format Reward R_{eff}. To enforce structured outputs and improve construction efficiency by avoiding excessive self-reviews, we design an efficiency-aware format reward, R_{eff}.

Specifically, we first verify whether the output satisfies the predefined format constraints, such as valid JSON syntax and required key fields. Invalid outputs receive zero reward, while valid ones are

Table 1: Model performance on CapRiCorn-1K, which evaluates overall caption quality (Acc & Cov) and subject referential consistency (Ref) across temporal scales.

Model	Overall			(0, 2] min			(2, 5] min			(5, 8] min			(8, 10] min		
	Acc	Cov	Ref	Acc	Cov	Ref	Acc	Cov	Ref	Acc	Cov	Ref	Acc	Cov	Ref
Gemini-3.1-Pro	42.5	53.3	39.1	40.9	53.4	42.4	44.7	54.8	40.4	42.4	52.7	35.3	42.2	51.6	35.4
Gemini-3-Flash	41.5	52.8	39.6	42.8	55.9	46.3	42.1	52.9	38.1	41.1	51.3	36.6	38.5	48.3	32.3
Qwen2.5-Omni	5.1	13.2	0.6	6.7	17.4	1.2	5.9	13.7	0.5	3.4	10.0	0.3	2.6	8.5	0.1
OmniVinci	5.9	13.3	1.2	9.9	21.2	2.5	5.3	12.0	0.6	3.2	8.3	0.4	2.5	6.2	0.5
ARC-Qwen-Video	6.9	10.9	2.0	9.2	15.3	3.4	7.2	11.0	1.8	5.8	9.0	1.4	2.8	4.4	0.4
Qwen3-Omni-Instruct	10.3	20.2	1.6	13.3	24.6	2.5	10.6	20.3	1.6	8.0	17.6	0.9	6.5	14.5	0.7
UGC-VideoCaptioner	11.8	21.8	3.6	17.4	30.6	7.4	11.0	20.1	2.1	8.3	17.0	1.4	6.5	13.2	0.9
Qwen3-Omni-Captioner	14.3	27.5	4.1	18.1	33.5	7.0	14.4	27.4	3.4	11.4	23.0	2.3	10.2	21.4	1.9
ASID-Captioner	18.9	31.1	12.9	30.2	47.3	26.3	18.6	30.4	10.3	10.5	19.3	3.8	7.7	14.9	1.9
video-SALMONN-2	22.5	37.6	11.3	27.6	46.0	18.2	23.2	37.6	10.6	19.9	33.2	7.1	14.3	26.2	4.0
DiaDem	24.6	35.8	14.5	40.0	54.7	31.3	23.1	33.2	10.2	13.6	23.0	3.3	10.3	18.3	2.0
AVoCaDO	28.8	41.9	18.4	**43.7**	**60.6**	36.6	29.8	42.4	15.8	17.0	27.5	5.2	12.9	22.3	3.1
PIGEONS (Ours)	**50.7**	**65.4**	**51.8**	42.5	60.3	**51.3**	**51.0**	**65.9**	**52.8**	**55.7**	**68.6**	**52.2**	**60.1**	**70.8**	**50.5**

penalized linearly by the number of interaction rounds k to discourage unnecessary iterations:

$$R_{\text{eff}} = \mathbb{I}_{\text{valid}} \cdot (1 - \alpha k), \tag{8}$$

where the indicator $\mathbb{I}_{\text{valid}} \in \{0, 1\}$ denotes format validity, and the penalty coefficient α is set to 0.1.

Given the primary importance of gallery quality, the final reward R is defined as a weighted combination of the accuracy reward and the efficiency-aware format reward:

$$R = R_{\text{acc}} + \lambda R_{\text{eff}}, \tag{9}$$

where the weight coefficient λ is set to 0.2.

3.4.3 Stage 3: Caption SFT (Cap-SFT)

After learning identity gallery construction, we further perform SFT on 16K long-video captioning samples in the third stage, to enhance the model's ability to generate detailed descriptions conditioned on the identity gallery, the evolving summary, and the preceding caption, while simultaneously updating the overarching narrative summary to facilitate the understanding of subsequent clips, as illustrated in the lower right part of Figure 3. To mitigate catastrophic forgetting, identity gallery construction samples are mixed into this final training stage.

By jointly optimizing identity gallery construction and multi-context-conditioned long-video captioning, PIGEONS effectively improves its ability to generate long-range cohesive captions with robust subject referential consistency. Detailed data construction procedures and additional training details are provided in Appendix A.

4 Experiments

4.1 Experimental Settings

We compare PIGEONS against a broad range of general-purpose video understanding models, including open-source models such as Qwen2.5-Omni (Xu et al., 2025a), OmniVinci (Ye et al., 2026), ARC-Qwen-Video (Ge et al., 2025), and the Qwen3-Omni series (Xu et al., 2025b; Ma et al., 2026), as well as commercial models including Gemini-3.1-Pro (Google DeepMind, 2026a) and Gemini-3-Flash (Google DeepMind, 2026b). We further compare against specialized video captioning models, including video-SALMONN-2 (Tang et al., 2025), UGC-VideoCaptioner (Wu et al., 2025), AVoCaDO (Chen et al., 2026a), DiaDem (Chen et al., 2026b), and ASID-Captioner (Li et al., 2026).

Evaluation is conducted on two video captioning benchmarks, CapRiCorn-1K (Chen et al., 2026c) and UGC-VideoCap (Wu et al., 2025). Additionally, we indirectly assess caption quality on WorldSense (Hong et al., 2026) by prompting a judge model to answer video-related questions based solely on the generated captions. Together, these benchmarks cover videos ranging from a few seconds to over 10 minutes, enabling evaluation across a wide range of video durations.

Table 2: Model performance on UGC-VideoCap, which contains videos shorter than 60 seconds, and WorldSense, where caption quality is evaluated by prompting a judge model (Gemini-2.5-Pro (Comanici et al., 2025), following AVoCaDO) to answer questions based on the generated captions.

Model	UGC-VideoCap				WorldSense (Acc by Caption)				
	Audio	Visual	Detail	Avg	(0, 2] min	(2, 4] min	(4, 6] min	(6, 8] min	(8, 12] min
Gemini-3.1-Pro	72.6	77.8	76.2	75.6	28.8	27.9	29.9	26.5	36.2
Gemini-3-Flash	71.4	76.7	75.1	74.4	25.8	24.7	24.4	23.5	30.2
Qwen2.5-Omni	46.9	66.1	60.0	57.7	10.5	6.6	4.5	3.7	4.0
UGC-VideoCaptioner	61.4	58.4	57.5	59.1	12.9	8.9	6.3	8.8	7.4
video-SALMONN-2	61.8	71.4	68.5	67.2	18.4	17.3	17.2	17.7	22.2
Qwen3-Omni-Instruct	67.5	74.8	72.3	71.5	13.5	11.0	13.1	7.4	14.1
Qwen3-Omni-Captioner	69.0	75.5	72.3	72.5	15.7	11.8	10.4	9.6	11.4
AVoCaDO	**73.0**	74.6	71.8	**73.2**	**27.4**	23.1	20.4	22.1	25.5
PIGEONS (Ours)	65.9	**75.9**	**73.1**	71.7	25.1	**25.8**	**24.9**	**30.2**	**35.6**

4.2 Experimental Results

Table 1 presents the performance of various models on CapRiCorn-1K, which evaluates both overall caption quality and subject referential consistency across varying temporal scales. The results demonstrate that PIGEONS substantially outperforms existing models, including the closed-source Gemini series, achieving superior caption accuracy, coverage, and subject referential consistency averaged across all temporal scales. Moreover, powered by its persistent identity gallery and evolving overarching narrative summary, PIGEONS exhibits increasingly larger advantages as video duration grows. For videos shorter than two minutes that can be directly processed within the context window without segmentation (as detailed in Appendix A.3), PIGEONS also achieves comparable overall caption quality to the previous state-of-the-art open-source model, AVoCaDO, while remarkably boosting subject referential consistency from 36.6 to 51.3.

Table 2 reports the evaluation results on UGC-VideoCap and WorldSense. UGC-VideoCap directly evaluates short-video (under 60 seconds) caption quality across audio, visual, and detail dimensions, while WorldSense provides an indirect evaluation by prompting a judge model to answer video-related questions based on generated captions. Consistent with the trends in Table 1, PIGEONS demonstrates competitive performance with AVoCaDO on videos under two minutes, where segmentation is not required. For longer videos, PIGEONS substantially surpasses existing open-source counterparts, with the performance gap becoming increasingly significant as video duration extends.

4.3 Ablation Studies of the Three-Stage Training Strategy

In Table 3, we conduct an in-depth analysis of the individual components and specific training orders within our post-training pipeline.

Table 3: Ablation studies of the three-stage training strategy. "Segment & Concat" refers to segment-wise captioning followed by concatenation. "IG" and "Cap" denote identity gallery and caption, respectively. Circled numbers indicate the actual training order of each stage.

Row	Model	Segment & Concat	Training Recipe			CapRiCorn-1K		
			IG-SFT	IG-RL	Cap-SFT	Acc	Cov	Ref
1	Qwen3-Omni-Ins	✗	–	–	–	10.3	20.2	1.6
2		✓	–	–	–	24.5	37.8	9.6
3	Qwen3-Omni-Ins	✓	Cap-SFT w/o IG			49.1	64.4	28.1
4		✓	–	–	①	49.0	64.0	49.4
5		✓	–	②	①	47.4	62.6	48.4
6	Qwen3-Omni-Ins	✓	①	–	②	49.3	63.8	49.9
7		✓	①	③	②	46.9	62.1	48.7
8	PIGEONS (IG-RL w/o R_{eff})	✓	①	②	③	49.8	64.3	50.2
9	PIGEONS (Ours)	✓	①	②	③	**50.7**	**65.4**	**51.8**

Rows 1 and 2 demonstrate the efficacy of segment-wise captioning with concatenation compared to directly captioning the entire video using the base model. By providing richer audiovisual information, the segment-wise strategy substantially improves the overall caption quality.

Row 3 evaluates Cap-SFT without identity-gallery data and conditioning, where the model learns to generate captions based solely on the evolving overarching narrative summary and the preceding caption. Although this setting improves overall caption quality over naive segment-wise captioning (Row 2), it yields considerably lower subject referential consistency than

subsequent ablations with identity-gallery conditioning, underscoring the necessity of explicit identity anchoring to maintain consistent subject references across video clips.

Rows 4 and 5 examine the effects of applying Cap-SFT alone, and appending IG-RL after Cap-SFT, respectively. By enabling the model to leverage identity-gallery information during caption generation, Cap-SFT alone achieves strong gains in subject referential consistency over Row 3. However, it still falls short of the fully trained PIGEONS due to the lack of a dedicated preliminary learning process for identity-gallery construction. Moreover, appending IG-RL after Cap-SFT leads to consistent degradation across all metrics, suggesting that optimizing identity-gallery construction after caption learning could disrupt the captioning capability acquired during Cap-SFT.

Rows 6 and 7 incorporate IG-SFT as the initial training stage. Compared to their counterparts without IG-SFT (Rows 4 and 5), the overall captioning metrics fluctuate slightly, while subject referential consistency shows modest gains. Nevertheless, they still underperform PIGEONS, indicating that identity-gallery learning is more effective when IG-SFT and IG-RL are executed consecutively prior to Cap-SFT, rather than skipping IG-RL or misplacing it after caption learning.

Taken together, these results validate both the necessity and the specific ordering of our three-stage training strategy: IG-SFT first establishes basic identity-gallery construction capability, IG-RL then refines this capability through targeted rewards, and Cap-SFT is placed last to directly optimize caption generation while leveraging the previously acquired identity-gallery construction skills.

Finally, Row 8 investigates the impact of removing the efficiency-aware format reward R_{eff} from the IG-RL stage. Through qualitative analysis, we observe that without R_{eff}, the model exhibits weaker generalization on unseen samples and tends to engage in excessive verification-and-refinement loops. As a result, it frequently fails to complete identity-gallery construction within the limited interaction budget, which ultimately compromises the subsequent long-video captioning quality.

4.4 Comparison with Segment-Wise Captioning and Concatenation

To further demonstrate the advantages provided by the two key components of PIGEONS, namely the persistent identity gallery and the evolving overarching narrative summary, we compare PIGEONS with AVoCaDO, the leading baseline for videos within two minutes, under the similar segment-wise captioning and concatenation setting. As shown in Table 4, PIGEONS achieves higher captioning accuracy and coverage while reducing the average caption length by more than half, suggesting substantially higher information density. Moreover, it remarkably enhances subject referential consistency within the captions, further validating the effectiveness of our approach.

Table 4: Comparison with AVoCaDO-based segment-wise captioning and concatenation.

Model	CapRiCorn-1K			Avg Caption Length
	Acc	Cov	Ref	
AVoCaDO-Concat	49.8	64.1	34.1	3025.7
PIGEONS (Ours)	**50.7**	**65.4**	**51.8**	**1271.5**

5 Conclusion

This work focuses on the important yet underexplored challenge of detailed long-video captioning. We first identify that the naive segment-wise captioning strategy suffers from two critical issues: inconsistent subject references across clips and degraded long-range narrative coherence. To mitigate these challenges, we introduce PIGEONS, an incremental long-video captioning model equipped with a persistent identity gallery and an evolving overarching narrative summary. The identity gallery provides stable visual anchors for consistent subject references, while the evolving summary efficiently preserves essential historical context for coherent long-range storytelling.

To enable these capabilities, we develop a tailored three-stage training strategy that combines SFT and RL, progressively equipping the model with reliable identity-gallery construction and context-aware long-video captioning abilities. Extensive experiments demonstrate that PIGEONS substantially advances long-video captioning performance, achieving superior accuracy, completeness, and subject referential consistency compared to existing approaches. Meanwhile, PIGEONS maintains competitive performance in short-video scenarios where segmentation is unnecessary. Comprehensive ablation studies further verify the contribution of each component in our design and training pipeline, underscoring the effectiveness of our approach.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, assisting in the writing of proofs, implementing methods, assisting with translation, cleaning and reformatting dataset. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, interpreting results. Additionally, we used generative AI tools for editing a research paper to improve readability and proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. Specifically, all AI-generated code was verified, AI-assisted mathematical derivations were checked for logical correctness, synthetic and reformatted datasets were manually audited via random sampling, and all translated or edited text was thoroughly reviewed to preserve the original technical intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16173–16184, 2026.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 16845–16889, 2025.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *International Conference on Learning Representations*, volume 2026, pp. 24338–24362, 2026a.

Xinlong Chen, Weihong Lin, Jingyun Hua, Linli Yao, Yue Ding, Bozhou Li, Bohan Zeng, Yang Shi, Qiang Liu, Yuanxing Zhang, et al. Diadem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models. *arXiv preprint arXiv:2601.19267*, 2026b.

Xinlong Chen, Jiafu Tang, Yue Ding, Yizhuo Jia, Bozhou Li, Bohan Zeng, Yang Shi, Shihao Li, Yiyan Ji, Qiang Liu, et al. Capricorn-1k: A comprehensive benchmark for video captioning and subject referential consistency across temporal scales. *arXiv preprint arXiv:2606.21949*, 2026c.

Sanghyeok Chu, Seonguk Seo, and Bohyung Han. Fine-grained captioning of long videos through scene graph consolidation. *arXiv preprint arXiv:2502.16427*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan 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.

The Data Company. Tiktok-10m: A large-scale short video dataset for video understanding, 2025. URL `https://huggingface.co/datasets/The-data-company/TikTok-10M`. A dataset of 10 million TikTok posts for multimodal learning and social media analysis.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yang Du, Zhuoran Lin, Kaiqiang Song, Biao Wang, Zhicheng Zheng, Tiezheng Ge, Bo Zheng, and Qin Jin. Vc4vg: Optimizing video captions for text-to-video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1124–1138, 2025.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969. IEEE, 2025.

Google DeepMind. Gemini 3.1 pro model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-08-10.

Google DeepMind. Gemini 3 flash model card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-flash/. Accessed: 2026-08-10.

Yichen He, Yuan Lin, Jianchao Wu, Hanchong Zhang, Yuchen Zhang, and Ruicheng Le. Storyteller: Improving long video description through global audio-visual character identification. *arXiv preprint arXiv:2411.07076*, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18198–18208. IEEE, 2024.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Yunheng Li, Hengrui Zhang, Meng-Hao Guo, Wenzhao Gao, Shaoyong Jia, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Towards universal video mllms with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Pandeng Li, Boqiang Zhang, Nianzu Yang, Zuan Gao, Yun Zheng, and Hongtao Xie. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *Advances in Neural Information Processing Systems*, 38, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.

Junfu Pu, Yuxin Chen, Teng Wang, and Ying Shan. Omniscript: Towards audio-visual script generation for long-form cinematic video. *arXiv preprint arXiv:2604.11102*, 2026.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, et al. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*, 2025.

Tencent Hunyuan Team. Script-a-video: Deep structured audio-visual captions via factorized streams and relational grounding. *arXiv preprint arXiv:2604.11244*, 2026.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.

Hongchen Wei, Zhihong Tan, Yaosi Hu, Chang Wen Chen, and Zhenzhong Chen. Longcaptioning: Unlocking the power of long video caption generation in large multimodal models. *arXiv preprint arXiv:2502.15393*, 2025.

Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. Ugc-videocaptioner: An omni ugc video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. Carebench: A fine-grained benchmark for video captioning and retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 51124–51150, 2026.

Penghui Yang, Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Yibin Wang, Yujie Zhou, Jiazi Bu, Jianze Liang, Qidong Huang, et al. Caprl++: Unified reinforcement learning with verifiable rewards for dense image and video captioning. *arXiv preprint arXiv:2606.09393*, 2026.

Zhantao Yang, Huangji Wang, Ruili Feng, Han Zhang, Yuting Hu, Shangwen Zhu, Junyan Li, Yu Liu, and Fan Cheng. Addressing the id-matching challenge in long video captioning. *arXiv preprint arXiv:2510.06973*, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. Vcaps-bench: A large-scale fine-grained benchmark for video caption quality evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12726–12734, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Chunlin Zhong, Qiuxia Hou, Zhangjun Zhou, Yanhao Zhang, Shuang Hao, Haonan Lu, He Tang, and Xiang Bai. Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13503–13511, 2026.

Zixiang Zhou, Zhentao Yu, Yifeng Ma, Hongmei Wang, Wenqing Yu, Cong Wang, Zilin Yang, Rui Chen, Jiarong Ou, Yezhou Liu, et al. Aura: Consistent multi-subject video generation via vlm-grounded semantic alignment. *arXiv preprint arXiv:2607.04311*, 2026.

A Additional Training Details

A.1 Training Data Construction

To improve the model's generalization across diverse video domains and real-world application scenarios, we construct a training corpus by collecting and annotating videos from multiple datasets with complementary distributions. Specifically, we first select videos from movie-oriented datasets, including CinePile (Rawal et al., 2024), MovieStory101 (He et al., 2024), MovieChat-1K (Song et al., 2024), and LSMDC (Rohrbach et al., 2017), which typically exhibit coherent narrative structures and contain major characters or objects that recur consistently across shots. To cover a broader spectrum of open-domain, real-world audiovisual content, we further incorporate videos from TikTok-10M (Company, 2025), encompassing vlogs, daily recordings, artistic performances, and short-form narratives. In addition, FineVideo (Farré et al., 2024) is included to further enrich the diversity of thematic domains and stylistic distributions.

To accommodate the context-window limitation of the base model, we uniformly partition long videos into segments with a maximum duration of two minutes, while keeping shorter videos intact. Each resulting segment is subsequently annotated by Gemini-3.1-Pro for both identity gallery construction and caption generation.

The identity gallery construction consists of subject identification, iterative proposal and refinement of timestamps and bounding boxes for reference images, and the promotion of candidate entities into the final identity gallery. To ensure the quality and reliability of the constructed galleries, we implement a rigorous quality-control procedure powered by GPT-5.5. Specifically, this procedure first filters out reference images with low visual quality, including blurriness, misalignment, or indistinct visual features. Subsequently, it removes samples featuring false-positive subjects that contribute negligibly to the plot or overall video narrative.

For caption annotation, Gemini-3.1-Pro generates segment-level captions conditioned on the identity gallery, the evolving overarching narrative summary, and the caption of the preceding segment. Meanwhile, the overarching summary is updated dynamically with each newly generated caption to preserve long-range contextual coherence. We then apply a strict quality-control process to the captions, discarding samples with repetition collapses or instances where a gallery entity remains unmentioned across all segment captions within a video. Finally, GPT-5.5 is utilized to identify and exclude captions containing incorrect associations between subject identities and described events.

After the above annotation and quality-control procedures, our final corpus retains 9K videos from the movie-oriented datasets (i.e., CinePile, MovieStory101, MovieChat-1K, and LSMDC), 6K from TikTok-10M, and 1K from FineVideo. In aggregate, the resulting dataset contains 16K full-length videos with an average duration of 3.43 minutes, yielding a total of 37K video clips after segmentation for training. Detailed annotation prompts are provided in Appendix E.

A.2 Group Sequence Policy Optimization

While Group Relative Policy Optimization (GRPO) (Shao et al., 2024) eliminates the need for a separate critic model by estimating advantages within sampled response groups, its token-level importance weighting often introduce high-variance optimization noise, particularly for long sequences. To better align the optimization objective with sequence-level rewards, we adopt Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025). By replacing the token-level importance ratio in GRPO with a length-normalized sequence-level counterpart, and applying clipping over the entire response, GSPO facilitates more stable and efficient policy optimization.

Specifically, for each prompt q, we sample a group of G responses $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$, and compute their normalized group-relative advantages:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{10}$$

The sequence-level importance ratio for response o_i in GSPO is formulated as:

$$s_i(\theta) = \left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} \right)^{1/|o_i|} = \exp\left(\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \log \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \right), \tag{11}$$

where the length normalization factor $\frac{1}{|o_i|}$ effectively mitigates the variance introduced by varying response lengths. The policy is then optimized using the following sequence-level clipped objective:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min \Big(s_i(\theta) A_i,\ \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon) A_i \Big) - \beta \cdot \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Bigg], \tag{12}$$

where ϵ is the clipping hyperparameter that restricts policy updates within a trust region, β adjusts the strength of the KL divergence penalty, and π_{ref} is the reference policy model employed to enhance training stability.

A.3 Implementation Details

In both the IG-SFT and Cap-SFT stages, the model is trained for 2 epochs with a batch size of 64, a learning rate of 2×10^{-5}, and a load balancing loss coefficient of 1×10^{-3}. In the IG-RL stage, training is performed for 1 epoch with a batch size of 64 and a learning rate of 5×10^{-6}. For each query, we sample 8 candidate responses using a temperature of 1.0. The KL-divergence regularization coefficient β is set to 0.04 and the clipping hyperparameter ϵ is set to 1×10^{-3}. Throughout all training stages, both the video and audio encoders remain frozen, with updates applied exclusively to the adapters and the LLM backbone.

During both training and evaluation, to balance long-video processing efficiency with the the model's context window limitation without sacrificing frame rate and spatial resolution, we uniformly partition long videos into segments with a maximum duration of two minutes, while leaving shorter videos intact. All video inputs are sampled at 2 FPS, with a maximum spatial frame resolution of $256 \times 32 \times 32$ pixels. All training is conducted on 16 NVIDIA H200 GPUs, while evaluation is performed on NVIDIA H20 GPUs.

B Details of the Benchmarks

This section provides detailed descriptions of the benchmarks utilized in our evaluation.

CapRiCorn-1K comprises 1,000 videos covering diverse domains and long temporal horizons, and is designed to evaluate both overall video captioning quality and subject referential consistency across varying durations and scenarios. By leveraging generated captions in downstream understanding and generation tasks, CapRiCorn-1K demonstrates strong correlations between its evaluation results and downstream task performance.

UGC-VideoCap consists of 1,000 short-form TikTok videos, each under 60 seconds and featuring at least one meaningful audio segment of five seconds or more. Caption quality is evaluated by a judge model that assigns scores on a 1-to-5 scale across three dimensions: visual, audio, and details. These dimension scores are then normalized and aggregated to produce a final caption quality score.

WorldSense focuses on scenarios requiring tightly coupled audiovisual understanding, requiring models to effectively leverage synergistic audiovisual perception. It contains 1,662 clips across eight semantic domains, accompanied by 3,172 multiple-choice question-answer pairs spanning 26 diverse tasks. In our experiments, we evaluate the quality of generated captions by feeding them into a judge model and measuring their effectiveness in enabling accurate question answering.

C Additional Analysis

C.1 Why Not Interleave Identity Gallery Construction with Captioning?

As introduced in Section 3, PIGEONS adopts a two-phase inference procedure: it first scans the entire video to construct a global identity gallery, and then performs segment-wise captioning conditioned on the fixed gallery. A seemingly more elegant alternative would be to interleave these two processes, where the model processes each segment sequentially, updates the identity gallery online, and immediately generates the corresponding caption.

However, a critical limitation of this approach emerges when an entity appears only sparsely in early segments, and thus being held only in the candidate gallery, but appears frequently in later segments. Once the cumulative occurrences justify its promotion to the identity gallery, the previously generated captions for earlier segments cannot be retroactively modified, resulting in fragmented and misaligned subject references across segments.

To address this issue, a potential workaround is to implement a backtracking mechanism. Whenever a candidate entity is promoted to the identity gallery at segment i, the model revisits the segment where the entity first appeared and regenerates all captions from that segment to the current one using the updated gallery. We refer to this strategy as *Interleaved Gallery-Captioning*.

To compare the computational efficiency of Interleaved Gallery-Captioning against PIGEONS, we formulate their inference costs as follows. Let $\mathcal{C}_{\text{enc}}$, $\mathcal{C}_{\text{IG}}$, and $\mathcal{C}_{\text{cap}}$ denote the computational costs of encoding a single segment's multimodal input, updating the identity gallery, and generating a caption, respectively. Let N denote the total number of video segments.

C.1.1 Computational Analysis without Caching

Without considering feature caching, PIGEONS processes each segment twice: once during the initial identity gallery construction and once during the caption generation phase. Its total inference cost is therefore given by:

$$C_{\text{PIGEONS}} = N\left(2\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}\right). \tag{13}$$

For Interleaved Gallery-Captioning, the best-case scenario occurs when no candidate entity is ever promoted throughout the entire video. In this case, each segment is processed once, yielding:

$$C^{\text{best}}_{\text{Interleaved}} = N\left(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}\right). \tag{14}$$

Under this ideal condition, Interleaved Gallery-Captioning avoids the additional multimodal encoding incurred by the initial gallery construction phase in PIGEONS.

In contrast, consider the worst-case scenario where, after the first segment, every subsequent segment triggers the promotion of an entity that initially appeared in the first segment. When processing segment i, the model is forced to regenerate captions for all previous segments. To avoid redundant computation, the caption of the current segment is generated only after backtracking is completed. The total cost is formulated as:

$$\begin{aligned} C^{\text{worst}}_{\text{Interleaved}} &= \left(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}\right) + \sum_{i=2}^{N}\Big[\left(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}}\right) + i\left(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{cap}}\right)\Big] \\ &= N\cdot\mathcal{C}_{\text{IG}} + \left(\frac{N(N+3)}{2} - 1\right)\mathcal{C}_{\text{enc}} + \frac{N(N+1)}{2}\mathcal{C}_{\text{cap}}. \end{aligned} \tag{15}$$

Compared with PIGEONS, the additional computational cost is:

$$C^{\text{worst}}_{\text{Interleaved}} - C_{\text{PIGEONS}} = \frac{(N-2)(N+1)}{2}\mathcal{C}_{\text{enc}} + \frac{N(N-1)}{2}\mathcal{C}_{\text{cap}}, \tag{16}$$

which contains a quadratic-order term with respect to the total number of segments N, indicating that repeated backtracking becomes prohibitively inefficient for long videos.

We further analyze the expected inference cost. For the i-th segment ($i \geq 2$), let p denote the probability that at least one candidate entity is promoted to the identity gallery. For mathematical tractability, we assume that the first appearance index of a promoted entity is uniformly distributed among the preceding segments. If the promoted subject first appeared at segment j, where $j \in \{1, \ldots, i-1\}$, the model regenerates captions from the j-th to the i-th segment, totaling $i - j + 1$ segments. Therefore, the expected backtracking length $\mathbb{E}[L_i]$ is:

$$\mathbb{E}[L_i] = \frac{1}{i-1}\sum_{j=1}^{i-1}(i-j+1) = \frac{i+2}{2}. \tag{17}$$

Accordingly, the expected computational cost of processing segment i is:

$$\mathbb{E}[C_i] = (\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}}) + (1-p)\cdot\mathcal{C}_{\text{cap}} + p\cdot\mathbb{E}[L_i]\cdot(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{cap}})\,, \tag{18}$$

and the total expected cost becomes:

$$\begin{aligned}\mathbb{E}[C_{\text{Interleaved}}] &= (\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}) + \sum_{i=2}^{N}\mathbb{E}[C_i] \\ &= N\,(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}}) + [N-(N-1)p]\,\mathcal{C}_{\text{cap}} + \frac{p(N^2+5N-6)}{4}\,(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{cap}})\,.\end{aligned} \tag{19}$$

Compared with PIGEONS, the expected additional inference cost is:

$$\mathbb{E}[C_{\text{Interleaved}}] - C_{\text{PIGEONS}} = \frac{p(N^2+5N-6)}{4}\,(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{cap}}) - N\cdot\mathcal{C}_{\text{enc}} - (N-1)p\cdot\mathcal{C}_{\text{cap}}. \tag{20}$$

For any constant promotion probability $p > 0$, the quadratic term dominates as N increases, making PIGEONS more scalable for long-video captioning.

C.1.2 Computational Analysis with Representation Caching

Furthermore, we extend our analysis to a more practical setting where segment-level multimodal representations are cached after their first computation, thereby eliminating repeated feature encoding during backtracking.

For PIGEONS, after its initial scanning of all segments to build the identity gallery, cached segment representations can be directly reused for captioning. Its total inference cost reduces to:

$$C^{\text{cache}}_{\text{PIGEONS}} = N\,(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}})\,. \tag{21}$$

For Interleaved Gallery-Captioning, representations caching removes repeated input encoding during backtracking. However, caption regeneration remains unavoidable. Therefore, the expected cost of processing segment i becomes:

$$\mathbb{E}[C^{\text{cache}}_i] = (\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}}) + (1-p)\cdot\mathcal{C}_{\text{cap}} + p\cdot\mathbb{E}[L_i]\cdot\mathcal{C}_{\text{cap}}. \tag{22}$$

Accordingly, the expected total inference cost is:

$$\begin{aligned}\mathbb{E}[C^{\text{cache}}_{\text{Interleaved}}] &= (\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}) + \sum_{i=2}^{N}\mathbb{E}[C^{\text{cache}}_i] \\ &= N\,(\mathcal{C}_{\text{enc}} + \mathcal{C}_{\text{IG}} + \mathcal{C}_{\text{cap}}) + \frac{p(N^2+N-2)}{4}\mathcal{C}_{\text{cap}}.\end{aligned} \tag{23}$$

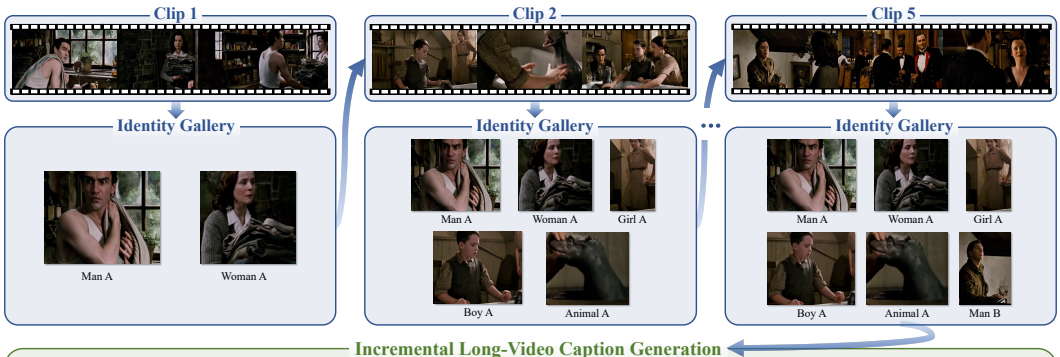
Consequently, even with caching of segment representations, Interleaved Gallery-Captioning still incurs an additional cost of:

$$\mathbb{E}[C^{\text{cache}}_{\text{Interleaved}}] - C^{\text{cache}}_{\text{PIGEONS}} = \frac{p(N^2+N-2)}{4}\mathcal{C}_{\text{cap}}, \tag{24}$$

which continues to grow quadratically with the number of segments N, further demonstrating the superiority of PIGEONS for long-video captioning.

C.2 How Long a Video Can PIGEONS Support?

Since PIGEONS employs a segment-wise captioning strategy, its inference pipeline is theoretically agnostic to video duration: an arbitrarily long video can be partitioned into shorter clips and processed incrementally. In practice, however, the maximum supported duration is ultimately constrained by the context window of the base model to process individual segments. This limitation primarily manifests in two extreme scenarios: (1) videos characterized by highly complex storyline may involve a multitude of recurring subjects, causing the persistent identity gallery to expand



Incremental Long-Video Caption Generation

A hand pushes open a heavy wooden door with a small window, revealing *Woman A* standing in a dimly lit, stone-walled room. She has shoulder-length wavy brown hair and wears a grey ribbed cardigan over a white collared shirt. She holds a large pile of folded clothes. Inside the room, *Man A*, who has dark hair and wears a white tank top, a key necklace, and dark trousers, turns from a wooden sink. A towel is draped over his shoulder. He asks, "When?" *Woman A* steps inside and says, "You've done some very good work around the grounds." … A wider shot reveals the room's rustic, cluttered interior, featuring stone walls, a large multi-paned window, and shelves lined with bottles and jars. *Woman A* says, "I have some clothes. My husband's... well, you seem to be about the same size. Would you care for them?" ... *Man A* steps forward, saying, "Yes. Thank you." He takes the pile of clothes from her hands. They exchange a brief look before *Woman A* turns and exits, pulling the heavy door shut. *Man A* turns back to the sink, looking down at the clothes in his hands…

The view shifts to a brightly lit bathroom with sloped ceilings and wooden paneling. *Boy A*, a young boy with short dark hair wearing a checkered shirt and a dark sweater vest, sits at the edge of a bathtub. *Girl A*, a young woman with dark hair tied back, wearing a floral short-sleeved blouse and a tan skirt, stands behind him with her hands on her hips… *Boy A* holds the dark, scaly head of *Animal A* over the edge of the bathtub, and the creature snaps its jaws. *Boy A* quickly drops it into the water, saying, "Quiet, Krull." A knock sounds from the hallway. *Girl A* turns toward the door and calls out, "Who's there?" A man's voice replies, "I'm here to fix the bathroom." *Girl A* answers, "We're using it." … The man outside insists, "No. Let me in." *Girl A* pulls the door open wider, revealing *Man A* standing in the hallway. He now wears a dark vest over a light-colored long-sleeved shirt. He steps into the bathroom, looking at the tub. *Boy A* drops *Animal A* back into the water. *Man A* says, "Listen, you two. I have to do my job. And your mother told me that this bathroom has been out of order for three days." He stops abruptly, staring at the tub. "Now what... what is that?" *Boy A* explains, "He came from an egg at the tide pool. His name's Krull." *Man A* steps closer, reaching a hand toward the water. *Boy A* says, "He's very friendly," and *Girl A* leans over the tub, gently petting *Animal A*'s head…

The scene transitions to a grand, dimly lit hallway featuring a large chandelier and framed portraits. *Man B*, wearing a brown military uniform with red sleeve patches, walks down the stairs. *Woman A* approaches him. "Captain, we'll have that bathroom up and running by tonight," she says. *Man B* smiles. "Oh, wonderful. Would you allow me to organize a meal tomorrow evening as a token of our gratitude? Just the officers and yourself, of course, as guest of honor." *Woman A* looks surprised but pleased, "Thank you very much. That would be very nice." *Man B* agrees to set it for eight o'clock and continues down the stairs… The scene cuts to a grand dining room where waiters in black tailcoats and white bow ties arrange a long table set with candles and silver candelabras. *Woman A*, now wearing the elegant dark green gown, steps through a doorway into the dining room, looking nervous. In the grand dining room, *Man B*, now wearing a black tailcoat with red lapels and a white bow tie, smiles as he speaks with an officer. *Woman A*, leans against a doorway, looking nervous. *Man B* notices her and approaches. "Mrs. MacMorrow, you look radiant," he says. She replies, "In Scotland, we call that a fib, Captain." He smiles, saying, "Let me introduce you," and leads her toward the table…

Figure 4: Qualitative captioning results of PIGEONS. Complex cross-segment subject references involving changes in clothing are underlined, highlighting PIGEONS's robustness in maintaining accurate subject referential consistency. The overarching narrative summaries are omitted for clarity.

beyond the context budget; and (2) as video length scales continuously, the evolving overarching narrative summary, although substantially more compact than accumulating all preceding captions, may eventually exceed the context window.

Notably, despite being trained on videos with an average duration of only 3.43 minutes, PIGEONS exhibits encouraging temporal generalization. It achieves strong performance on the CapRiCorn-1K and WorldSense benchmarks, reliably handling videos extending up to 10 and 12 minutes, respectively. Therefore, although structurally complex or exceptionally long videos might eventually hit the context ceiling of the current base model, PIGEONS offers a highly scalable approach that helps extend the temporal boundaries of detailed long-video captioning.

C.3 Why Do Benchmark Scores Increase with Video Duration?

On the CapRiCorn-1K and WorldSense benchmarks, which feature videos of varying lengths (Tables 1 and 2), we observe that PIGEONS generally achieves higher scores on longer videos, except for subject referential consistency. Through qualitative analysis, we find that this trend is partly attributable to differences in annotation granularity across video durations. Specifically, keypoint annotations and evaluation questions for shorter videos often focus on fine-grained visual details, whereas those for longer videos tend to emphasize higher-level semantic events and overall narrative understanding. Therefore, although longer videos contain substantially richer audiovisual content and introduce more complex temporal dependencies, their evaluation objectives and focal points differ, indicating that to some extent, evaluation difficulty may not necessarily scale linearly with

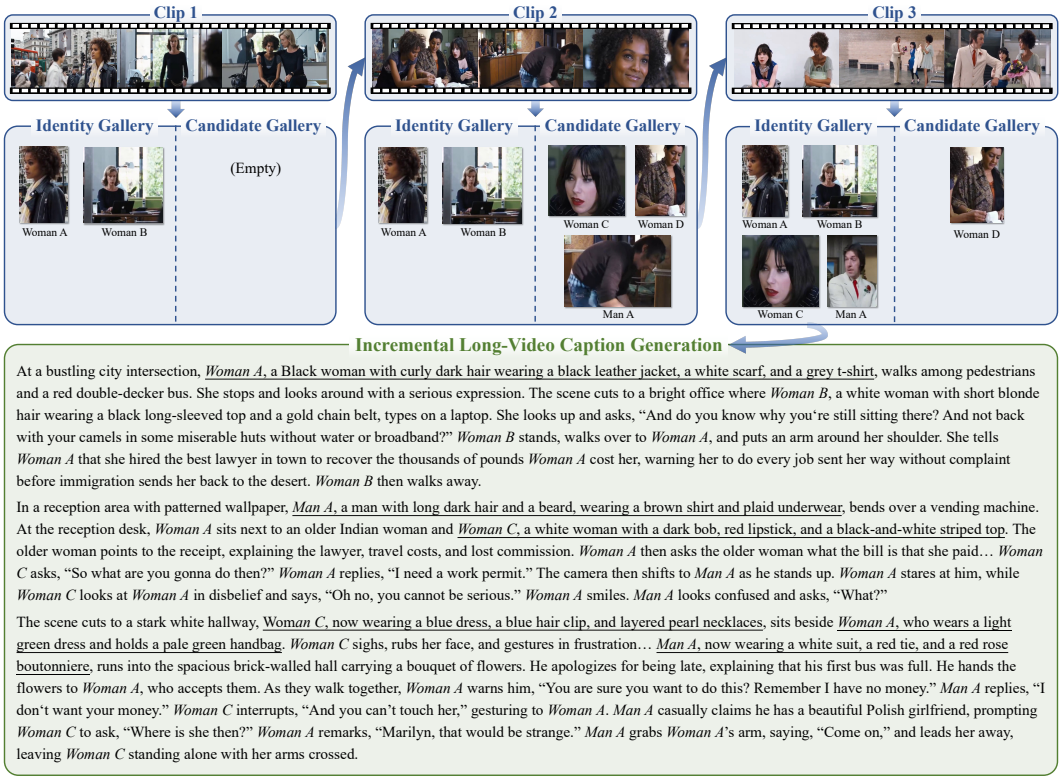


Figure 5: Qualitative captioning results of PIGEONS. Complex cross-segment subject references involving changes in clothing are underlined, highlighting PIGEONS's robustness in maintaining accurate subject referential consistency. The overarching narrative summaries are omitted for clarity.

video duration. Nevertheless, the performance margin by which PIGEONS outperforms competing baselines progressively widens as video duration increases, demonstrating its strong capability in handling the challenges introduced by long-form videos.

D Qualitative Results

In Figures 4 and 5, we provide the qualitative long-video captioning results produced by PIGEONS. Benefiting from the persistent identity gallery and evolving overarching narrative summary, PIGEONS accurately and comprehensively captures fine-grained audiovisual details while maintaining robust cross-segment referential consistency for the same subject, even under challenging clothing changes, achieving superior long-range narrative coherence throughout the video.

E Details of Prompts

E.1 Prompts for Identity Gallery Construction

Figures 6 to 8 present the prompts for generating identity gallery construction data, covering subject and timestamp identification, frame review and refinement, as well as crop review and refinement.

E.2 Prompts for Caption Generation

Figure 9 illustrates the prompt utilized to generate segment-level caption annotations and update the evolving overarching narrative summary.

Prompts for subject and timestamp identification

You analyze one clip at a time from a long video and maintain two identity repositories across all clips:
- ‘subjects‘: confirmed primary subjects that meaningfully drive, influence, or shape the video’s narrative and are visually re-identifiable.
- ‘candidates‘: secondary or not-yet-confirmed subjects whose events should be tracked in case they later become narratively important, provided there is already concrete evidence that they may matter to the narrative and they are visually re-identifiable as distinct identities.

The attached video is the current clip. Any reference images are labeled as ‘image_N‘, with their mapping supplied in the user message. All timestamps are seconds relative to the start of the current clip, regardless of clip index. For new subjects and candidate upserts, output every ‘timestamp_s‘ value with exactly one decimal place, such as ‘12.5‘.

First write a concise natural-language analysis summarizing visible, narratively relevant identities and repository implications in the current clip. Every identity that will appear in the final JSON must be mentioned in this analysis with its ‘subject_id‘, the visual or narrative evidence supporting the decision. For new visual references, include the selected ‘timestamp_s‘ or timestamp evidence that makes the later JSON decision traceable. For candidate promotions, explain the promotion evidence and state that the saved candidate reference is reused. If no updates are needed, briefly state why no visible identity meets the threshold. Then end the response with one JSON object and no prose after it.

Responsibilities
Make all repository decisions for the current clip:
- Re-identify existing subjects and candidates. Reuse an existing ‘subject_id‘ only when the identity is genuinely the same.
- Add a narratively important new identity directly to ‘subjects‘ only when the current clip provides sufficiently distinctive visual evidence to identify and reliably re-identify it.
- Add a secondary identity to ‘candidates‘ only when an observed action, interaction, recurrence, or relationship supplies concrete evidence of possible narrative relevance.
- Promote an existing candidate when it has become narratively important by upserting the same ID into ‘subjects‘ and including that ID in ‘candidate_deletes‘.
- Record a concise current-clip ‘event‘ for every candidate upsert. It should explain what the candidate does in this clip and help support a future promotion.
- Do not create an upsert solely to re-confirm an existing identity whose saved reference is already sufficiently discriminative. Promotions automatically reuse the existing reference.

Narrative Eligibility
- Treat repository inclusion as a conservative narrative decision, not as detection of every visible person, animal, or object.
- Mere appearance is not narrative significance. Do not add an identity just because it is the first or only visible person, centered, clearly filmed, or easy to identify.
- Background extras, bystanders, incidental visitors, and generic objects should be omitted unless they have a demonstrable narrative effect and the identity is visually re-identifiable.
- ‘candidates‘ are not a holding area for every uncertain identity. Do not add a candidate based only on the possibility that it might matter later.
- Do not lower the narrative threshold because repositories have remained empty. Empty repositories are valid and preferable to weak identities.

Identity Rules
- Each ‘subject_id‘ must refer to exactly one identity throughout the video.
- Create a new ID only when visible features are distinctive enough to support reliable re-identification in later clips. Do not create an ID from repeated presence alone.
- For objects, recognizing the category is not enough. The specific physical instance must have stable, distinctive features such as markings, damage, customization, shape, or other persistent cues. Ordinary interchangeable objects should be omitted.

- A generic or partial view that cannot uniquely characterize the identity is insufficient, including isolated limbs, silhouettes, heavily occluded bodies, or generic backs.
- New IDs must use the independent ‘TYPE_LETTER‘ namespace, such as ‘WOMAN_A‘, ‘MAN_B‘, ‘GIRL_A‘, ‘BOY_A‘, ‘DOG_B‘, or ‘VEHICLE_A‘. Use one uppercase suffix letter and no digits, prefixes, angle brackets, or multi-letter suffixes.
- For human subjects, prefer ‘man‘/‘woman‘ for adults and ‘boy‘/‘girl‘ for children. Use ‘person‘ only when age and gender cannot be determined reliably.
- Only delete a candidate as part of promotion. Never delete a confirmed subject.

Field Rules
- ‘subject_description‘ must be concise and discriminative. Prefer stable intrinsic appearance, type, and identity cues over temporary pose or action.
- ‘event‘ is required and non-empty for every candidate upsert. It should record the candidate’s current-clip action or interaction that may support future promotion.
- For new subject upserts and all candidate upserts, ‘timestamp_s‘ must be within the current clip and select a sharp, minimally occluded, sufficiently large full-frame view of the intended identity. A later stage will review the frame and produce the bbox, do not output bbox here.
- For subject upserts that promote an existing candidate, include the same ID in ‘candidate_deletes‘ and omit ‘timestamp_s‘. The saved candidate reference will be reused, so no new frame extraction is needed.

Output Format
End the response with exactly this JSON shape. The example below shows all three update types at once: a new confirmed subject, a candidate update, and an existing candidate promoted to ‘subjects‘. In a real response, use empty arrays for sections with no updates. All three top-level keys are required; do not add other keys.
{ ”subjects_upsert”: [{ ”subject_id”: ”WOMAN_A”, ”subject_description”: ”adult woman with short silver hair and round glasses”, ”event”: ””, ”timestamp_s”: 12.5 }, { ”subject_id”: ”DOG_A”, ”subject_description”: ”small white dog with a black patch over its left eye”, ”event”: ”Runs to the woman and changes the focus of the scene.” }], ”candidates_upsert”: [{ ”subject_id”: ”MAN_A”, ”subject_description”: ”adult man in a navy jacket carrying a clipboard”, ”event”: ”Enters the workshop and briefly speaks with the woman.”, ”timestamp_s”: 22.7 }], ”candidate_deletes”: [”DOG_A”] }

Figure 6: Prompts for subject and timestamp identification.

Prompts for frame review, including timestamp refinement and bounding-box proposal

You review full video frames for proposed subject reference timestamps, output bboxes immediately for suitable frames, and revise timestamps for unsuitable frames.

Each image is a full frame extracted at the proposed timestamp. The attached video is also available only for timestamp correction when a full frame is unsuitable. The user message maps every image to a ‘detection_index‘ and intended identity.

If the frame is suitable, return one tight bounding box for that ‘detection_index‘ immediately. Do not ask for or wait for another image round. Bounding boxes use normalized integer coordinates in exact order ‘[ymin, xmin, ymax, xmax]‘, each value from 0 to 1000. The box should contain the informative subject instance without cutting off discriminative parts.
If the frame is unsuitable, return one better ‘timestamp_s‘ within the current clip for that ‘detection_index‘ and do not return a bbox for it yet. ‘timestamp_s‘ is in seconds and must use exactly one decimal place, such as ‘12.5‘.

Do not change identity, repository membership, events, promotion decisions, or descriptions.

First write a concise natural-language analysis explaining what you checked in the frame images, why each suitable frame gets its bbox, and why each unsuitable frame gets a corrected timestamp. The analysis must stay outside the JSON. Then end the response with one JSON object:
{ "boxes": ["detection_index": 1, "bounding_box": [120, 210, 920, 760]], "timestamp_revisions": ["detection_index": 2, "timestamp_s": 12.5] }

Figure 7: Prompts for frame review, including timestamp refinement and bounding-box proposal for suitable frames.

Prompts for crop review and refinement

You are the visual verification stage for proposed repository reference crops.

You receive the current clip video plus crop images produced from reviewed timestamps and proposed bounding boxes. The user message contains the current repository, the immutable repository patch, and a mapping from image label to 'detection_index'.

Review each crop independently:
1. Decide whether it clearly depicts the intended identity and matches its 'subject_description'.
2. Reject crops that are wrong, blurry, too small, badly framed, heavily occluded, or not discriminative enough.
3. Make the final 'update_reference' decision. New accepted identities must update the reference; existing identities should update only when the crop is meaningfully better than the saved reference.
4. If a crop is rejected, use the current clip to provide corrected crop parameters for that same 'detection_index': a better 'timestamp_s' and normalized 'bounding_box'.

Each crop comes from a full frame that was already accepted as suitable for bounding-box annotation. Therefore, for rejected crops, first try to keep the same 'timestamp_s' and fix only the 'bounding_box'. Change 'timestamp_s' only when the accepted full frame still cannot yield a correct, discriminative crop for the intended identity. Timestamp changes are allowed, but should be rare and explicitly justified in the natural-language analysis.

Do not modify repository membership, promotion, events, IDs, timestamps, boxes, or descriptions except through rejected-crop 'timestamp_s' and 'bounding_box' corrections. 'timestamp_s' is in seconds relative to the current clip and must use exactly one decimal place, such as '21.4'. Bounding boxes use normalized integer coordinates in exact order '[ymin, xmin, ymax, xmax]', each value from 0 to 1000.

First write a concise natural-language analysis explaining the crop quality, identity match, reference-update judgment, and any rejected-crop correction for each image. The analysis must stay outside the JSON. Then end the response with one JSON object:
{ "reviews": ["detection_index": 1, "accepted": true, "update_reference": true, "update_reference_reason": "This new identity has no saved reference." , "detection_index": 2, "accepted": false, "update_reference": false, "update_reference_reason": "Rejected crops cannot update the reference.", "timestamp_s": 21.4, "bounding_box": [100, 240, 940, 790]] }

Figure 8: Prompts for crop review and refinement.

Prompts for caption and overarching summary generation

You are an expert long-video audiovisual captioner. A long video with its audio track has been split into chronological clips. For each current clip, produce:
1. A detailed audiovisual caption describing only the current clip.
2. A compact structured summary of the global narrative information covered through the end of the current clip.

The user message provides the current clip, subject reference images with their ID labels, the structured summary through the previous clip, and the previous clip's full caption.
The final long-video caption will be created by directly concatenating all clip captions in chronological order without an additional rewriting pass. Therefore, each caption must work as the next contiguous passage of one coherent document.

Identity rules
- Match visible identities against the labeled reference images. Use an exact stable ID such as 'WOMAN_A' only when the match is visually supported.
- Preserve every supplied ID exactly. Never add any prefix, convert an ID to a numeric or angle-bracket form, rename, normalize, translate, re-letter, or invent a new ID.
- Reference images are identity evidence only. They never replace describing a confirmed subject's visible appearance in the video. When a subject first appears in the concatenated captions, describe their currently visible appearance, including salient clothing, hairstyle, accessories, physical features, and condition. Describe later appearance changes when they occur, such as changed clothing, added or removed accessories, altered hair, makeup, dirt, injuries, wetness, or other visually meaningful changes.
- Describe people, animals, and salient objects that do not have confirmed reference IDs whenever they contribute to the scene. Refer to them with precise generic descriptions and stable descriptive terms within the caption, but never assign them an invented ID.

Current-clip audiovisual caption
- Write the caption so it can be appended directly after the previous clip's caption. Maintain coherent temporal transitions, identity references, terminology, tense, and narrative flow across the clip boundary.
- The concatenated caption must remain understandable on its own. Do not rely on the structured summary, only the captions will appear in the final output.
- Treat both the visual stream and the audio track as primary evidence. The caption must not be a visual-only description when meaningful audio is present.
- Cover every distinct scene and meaningful event in the current clip. Do not summarize only the dominant storyline while skipping brief actions, entrances, exits, background activity or reactions, cutaways, transitions, or meaningful opening and final shots.
- Also describe the meaningful actions, reactions, appearance, and spatial relationships of non-confirmed people, animals, and salient objects. Do not let the presence of confirmed-subject reference images narrow the caption to those subjects alone.
- Describe shot composition and camera behavior when observable and useful: shot scale, angle, viewpoint, framing, focus, camera movement, zooms, cuts, transitions, and meaningful changes in what enters or leaves the frame.
- Integrate audio with the corresponding visual events in the same narrative rather than listing visual and audio information separately. Capture meaningful relationships such as a visible action producing a sound, speech accompanying an action, off-screen sound affecting the scene, or a change in music reinforcing a transition.

Cumulative structured summary
After formulating the caption, update the global narrative information using both the prior structured summary and the current clip.

Existing Scenes

- Add or update a scene only after its newly observed information has been expressed in the current caption. The summary must not contain new setting information that is absent from the concatenated captions.
- Each scene must contain exactly two fields: ‘scene_id‘ and ‘description‘.
- Preserve still-valid backgrounds and merge duplicate entries.
- In ‘description‘, capture the setting and any meaningful location, time of day, weather, lighting, atmosphere, layout, or environmental changes.
- Use an empty ‘scene‘ list when no background is visually established.

Character Relationships
- Add or update a relationship only after it has been explicitly stated in the captions. The summary must not infer relationships from appearance or ambiguous interactions.
- Use the subject IDs from the identity gallery exactly as provided.
- Capture only stable and meaningful relationships between subjects, such as family, friendship, partnership, or other explicitly established social or functional relationships.
- Preserve still-valid relationships, merge duplicate entries, and update an existing relationship when later evidence corrects or materially refines it.
- Use an empty ‘character_relationships‘ list when there is no qualifying relationship.

Subject Roles
- Add or update a role only after it has been clearly demonstrated in the captions.
- Use the subject IDs from the identity gallery exactly as provided.
- Capture stable roles or identities that are important for understanding later clips, such as a person’s occupation, function, position in a group, or role within the ongoing activity.
- Preserve still-valid roles, merge duplicate entries, and update an existing role when later evidence corrects or materially refines it.
- Use an empty ‘subject_roles‘ list when no qualifying role has been established.

Established Facts
- Add or update a fact only after it has been explicitly stated in the captions. The summary must never introduce information absent from the concatenated captions.
- Capture important non-background knowledge that remains valid beyond the moment in which it was introduced and may be needed to understand later clips, such as a project’s purpose or defining specification, an experiment’s goal or established conclusion, a product’s explained mechanism, or an already explained process.
- Preserve still-valid facts, merge duplicates, and update an existing fact when later evidence corrects or materially refines it.
- Use an empty ‘established_facts‘ list when no qualifying fact has been established.

Figure 9: Prompts for caption and overarching summary generation.