HowToNarrate: A General-Domain Benchmark for Synchronized Video Narration with External Knowledge

Xueyan Wang[1*], **Dingyi Yang**[2*], **Qin Jin**[1†]
[1]Renmin University of China
[2]Nanyang Technological University
wangxy1117@ruc.edu.cn, dingyi.yang@ntu.edu.sg, qjin@ruc.edu.cn

Abstract

We present ***HowToNarrate***, the first general-domain benchmark for Synchronized Video Narration. The benchmark contains 3.2K videos across seven domains, segmented into 37.5K clips with aligned narrations and associated external knowledge. Effective narration requires models to *understand visual scenes*, incorporate *relevant knowledge*, and produce *coherent*, *length-appropriate* descriptions. We systematically benchmark current Multimodal LLMs (MLLMs) on these abilities. Our analysis shows that existing MLLMs overemphasize knowledge retrieval while largely neglecting prior context (receiving less than 10% attention). Moreover, they often conflate narration context with external knowledge, leading to redundancy and incoherence. To mitigate these issues, we propose VideoNarrationAgent, a multi-agent framework that combines context compression, knowledge retrieval, and narration generation. Experiments demonstrate that our method significantly improves MLLM performance. Furthermore, instruction tuning on HowToNarrate enhances both context-awareness and length control, boosting Qwen2.5-VL's score from 25 to 84. Our dataset and codes are released at https://github.com/wangxueyan666/HowToNarrate.

1 Introduction

Traditional video-to-text tasks, such as video captioning (Xu et al., 2016; Caba Heilbron et al., 2015), aim to generate descriptions for visual content, focusing primarily on recognizing and labeling what is directly observable. For instance, given a scene showing a camel's hump, a caption might simply state: "A camel with a large hump". While such descriptions are useful for summarization or retrieval, they offer limited depth and context. In contrast, video narration or storytelling (Li et al., 2019) goes a step further by creating a coherent narrative that not only describes what is seen but also connects events over time. More importantly, narration often requires external knowledge to be informative and persuasive. Continuing the camel example, an informative narration might explain: "The camel's hump stores fat, allowing it to survive long journeys in the desert."—a statement that goes beyond what is visually present. This type of knowledge-enriched narration is critical for applications such as educational content, documentaries, and instructional videos.

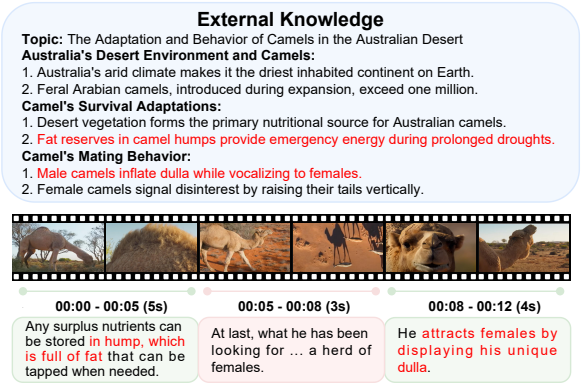


Figure 1: A case from our proposed dataset HowToNarrate. Each video contains sequential visual scenes paired with corresponding narrations. Each narration should relate to the visual content, incorporate relevant knowledge, have appropriate text length fitting the time duration, and maintain coherence with preceding narrations.

Moreover, to support real-world use cases like voiceover generation, video narration must satisfy synchronization constraints: the narration must match the timing and duration of each visual scene while preserving fluency and coherence. These requirements—knowledge enrichment, temporal alignment, and narrative consistency—define the emerging task of **Synchronized Video Narration** (Yang et al., 2024).

Existing efforts in video narration only partially addresses these challenges. Some focus solely

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39110–39131

on visual relevance, overlooking external knowledge and length control (Zhang et al., 2024a; Bhattacharya et al., 2023), while others are restricted to narrow domains such as advertising or entertainment (Yang et al., 2024; Sun et al., 2024). To address these limitations, we present **HowToNarrate**, the first large-scale, general-domain dataset for synchronized video narration. As illustrated in Figure 1, each HowToNarrate sample consists of: 1) a sequence of visual clips, 2) corresponding human-authored, time-aligned narrations, and 3) carefully curated external knowledge that supports or enhances the narration.[1] All annotations are manually verified for accuracy and consistency. The dataset consists of 3.2K videos segmented into 37.5K clips, totaling 156.2 hours, across diverse domains such as documentaries, popular science, handicrafts, and cooking. On average, each video contains 12 clips (3 minutes total), requiring models to handle both local and long-range coherence.

Using HowToNarrate, we thoroughly benchmark state-of-the-art Multimodal LLMs (MLLMs) across five key dimensions: *Visual Relevance, Fluency, Coherence, Word Count Accuracy* and *Knowledge Utilization* (both relevance and richness). Our evaluation, shown in Figure 2 and Table 3, reveals three **core limitations** in current MLLMs: 1) shallow or repetitive knowledge retrieval, 2) underuse of narration history, and 3) difficulty integrating visual, textual, and knowledge sources into coherent outputs. These findings highlight pressing challenges for advancing synchronized video narration.

To address these challenges, we propose **VideoNarrationAgent**, a lightweight yet effective multi-agent framework composed of three modules: *Narration Context Compression* to summarize prior narrations, *Knowledge Retrieval* to provide supporting information, and *Narration Generation* to produce temporally aligned, coherent output. Collaboratively, these modules guide MLLMs toward producing higher-quality narrations, with further gains achieved through instruction tuning on HowToNarrate.

Our contributions are threefold: 1) We introduce HowToNarrate, the first general-domain synchronized video narration benchmark with fine-grained narrations and aligned external knowledge. 2) We conduct a comprehensive evaluation of current MLLMs, revealing key limitations in knowledge integration and contextual modeling. 3) We propose VideoNarrationAgent, a multi-agent framework that significantly improves MLLM performance across all evaluation dimensions, offering practical insights for future research.

[1]E.g., a scene showing camel humps is enriched with background knowledge on their function in desert survival.

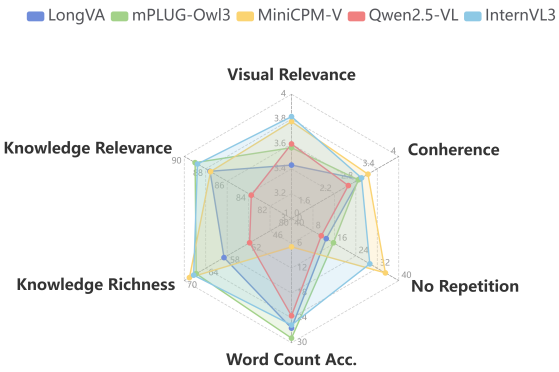


Figure 2: Performance of current MLLMs on synchronized video narration across multiple dimensions.

2 HowToNarrate Dataset

Existing video narration datasets suffer from three main shortcomings: 1) they lack fine-grained, synchronized narrations suitable for voiceover generation; 2) they are largely confined to narrow domains such as advertising and entertainment; and 3) they omit background information or external knowledge relevant to the video content. To overcome these limitations, we introduce **HowToNarrate**, the first general-domain synchronized video narration dataset. Each video is annotated with fine-grained, time-aligned narrations and enriched with relevant external knowledge (Figure 1). More data cases are provided in Appendix A.2.

2.1 Dataset Construction

Data Source. We collect raw narration videos from YouTube across diverse domains, including documentary, handcraft, popular science, cooking, travel, and tutorial. To ensure high quality, we source videos from reputable channels such as BBC Earth, WikiHow, Top Travel, and others. In total, we curate 3,214 videos, which are segmented into 37,522 video clips. For each video, we also download its title and category metadata.

Scene Segmentation and Narration Annotation. We segment videos into sequential clips based on frame-wise visual differences using a thresholding method (Huang et al., 2020). To obtain synchronized narrations, we apply a state-of-the-art Automatic Speech Recognition (ASR) model (Cao

et al., 2012). Since ASR outputs inevitably contain transcription errors, we further refine them using DeepSeek-V2.5 (Liu et al., 2024), as detailed in Appendix A.1.1. This refinement reduces the sentence-level error rate from 15.3% to 6.7%, substantially lowering the human correction workload.

External Knowledge Construction. Synchronized narrations extend beyond surface-level video descriptions by integrating relevant external knowledge to create informative and engaging narratives. To build the accompanying knowledge base, we adopt a two-step enrichment process[2]. First, we extract key knowledge points from the ground-truth (GT) human-written narrations. Second, we expand these points using DeepSeek's world knowledge to generate additional relevant information. This approach ensures that no GT knowledge is omitted while also providing models with a broader knowledge foundation for improved generalization. The annotation and purpose of external knowledge are detailed in Appendix A.1.2.

Quality Control. During data source selection, we manually filter out videos whose visual content has no correlation with the narration text. We ensure that the narration either directly describes the visual content or extends it in a relevant manner. For the LLM-refined ASR results, we manually correct residual typos and remove unrelated text elements such as channel information. In addition, annotators strictly review external knowledge with reference to the ASR results, ensuring that all entries in our external knowledge base are presented in a concise, abstract style while fully covering the professional information contained in the narration but not observable from the visual content. The inter-annotator agreement, measured by Pearson correlation on data qualification judgments, reached 0.62, indicating reasonable consistency among annotators.

2.2 Statistics and Analysis

To the best of our knowledge, HowToNarrate is the first general-domain benchmark for Synchronized Video Narration. As shown in Figure 3, it covers multiple domains, including documentary, handcraft, popular science, cooking, travel, and tutorial. HowToNarrate comprises 3,214 videos with an average duration of approximately 3 minutes. These videos contain 37,522 clips that range from 1 to 63 seconds, with an average length of 12 seconds. Figures 4 illustrate the distribution of video duration and the number of clips per video. Notably, around 29.1% of videos contain at least 20 clips, underscoring the challenge of generating long, coherent narrations. Furthermore, the task requires retrieving the most relevant visually grounded knowledge from an informative knowledge base (averaging 224 words per video), which increases complexity.

As shown in Table 1, videos in HowToNarrate have substantially longer average narration lengths compared to existing video narration datasets. This is primarily due to the requirement for synchronized narration and the extended duration of our videos. Additionally, we annotate external knowledge to support narration creation. Unlike knowledge in the advertising domain (Yang et al., 2024), our knowledge base is more diverse and unstructured, introducing further challenges for models.

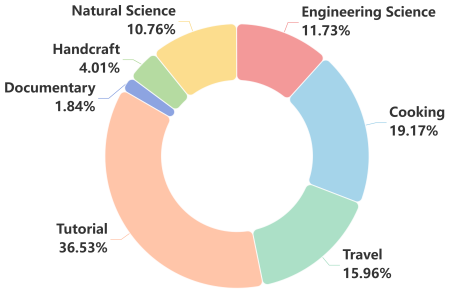


Figure 3: Video domain distribution in HowToNarrate.

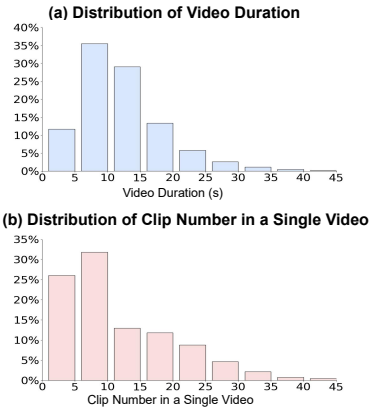


Figure 4: Distribution of video durations and clip counts per video in HowToNarrate.

3 Task and Evaluation

3.1 Task Definition

We focus on the synchronized video narration task, where the goal is to generate coherent, informative,

[2]To increase the challenge, constructed knowledge includes knowledge from GT and knowledge used to provide potentially useful information for generation.

	Domain	Input Modality	Synchronized Narrations	Video num.	Total video len.	Avg. video len.	Avg. clips (per video)	Avg. text len. (per video)
Video Storytelling (Li et al., 2019)	Open	V	-	105	22.0 h	12m 35s	13.5	162.6
Shot2Story20K (Han et al., 2023)	Open	V	-	20 k	94.4 h	17 s	4.0	71.2
VideoStory (Unreleased) (Gella et al., 2018)	Social Media	V	-	20 k	288.9 h	1m 10s	6.1	62.2
Video Persuasion (Bhattacharya et al., 2023)	Advertisement	V	-	1 k	9.2 h	33 s	-	-
E-SyncVidStory (Yang et al., 2024)	Advertisement	V + K	✓	6 k	65.0 h	39 s	6.9	194.1
HowToNarrate (Ours)	Open	V + K	✓	3 k	156.2 h	2m 55s	12.0	303.7

Table 1: A comparison between HowToNarrate and existing video narration datasets. In the Input Modality column, V represents video and K represents knowledge.

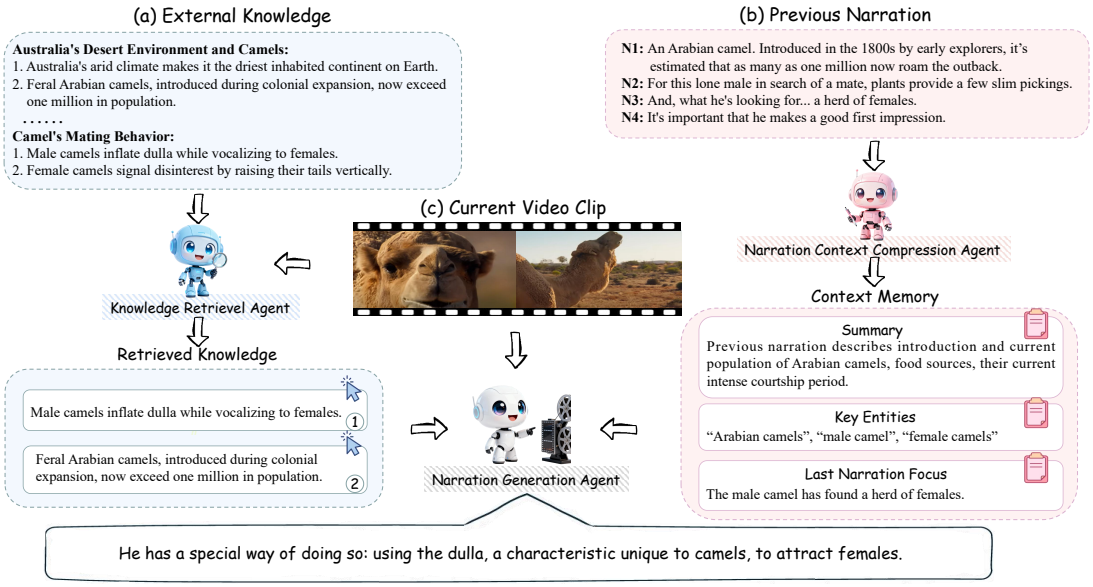


Figure 5: We design three different agents (§4): Narration Context Compression Agent, Knowledge Retrieval Agent, and Narration Generation Agent.

and temporally aligned narrations for video clips. Given a video V segmented into sequential clips $\{v_1, v_2, ..., v_n\}$ and a set of associated knowledge items $\mathcal{K} = \{k_1, k_2, ..., k_m\}$, the goal of synchronized video narration is to generate coherent narrations $\mathbf{S} = \{s_1, s_2, ..., s_n\}$, where each narration chunk s_i corresponds to clip v_i and incorporates relevant knowledge k_j that the clip is intended to convey. Each s_i is constrained to a target word range r_i.

3.2 Reference-based Evaluation

Following traditional video-to-text evaluation, we use reference-based metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015). These metrics measure the textual similarity between the generated narrations and the ground-truth.

3.3 Multi-dimensional Evaluation

Visual Relevance. We use EMScore (Shi et al., 2022) to measure how well the narration aligns with video content. EMScore only considers direct visual relevance, not logical relevance. However, video narration typically includes logical imagination and relevant knowledge. Therefore, we apply the advanced MLLM Gemini-2.5-Pro (Comanici et al., 2025)[3] to score Visual Relevance on a scale from 1-5, as detailed in Appendix A.3. Due to the inconsistency of MLLM-evaluation (Stureborg et al., 2024), we compute Vis_{rel} by averaging results over three independent runs. Additionally, we evaluate using other powerful MLLMs and calculate their Pearson correlations. As shown in Appendix A.4, the high correlations confirm the reliability of our MLLM-based evaluation.

Knowledge Utilization. Effective video narrations should both incorporate relevant knowledge for persuasiveness, measured by Know_{rel}, and cover a diverse set of knowledge items, measured by $\text{Know}_{\text{rich}}$. Following Yang et al. (2024), we compute these metrics at the sentence level. Specifically, Know_{rel} is the average of the maximum similarity between each narration and all knowledge items, while $\text{Know}_{\text{rich}}$ is the ratio of knowledge items covered (defined as those with similarity >

[3]Gemini-2.5-Pro is not included in our evaluated models.

0.85) relative to the total number of video clips. Formally, given video narrations $\mathcal{S} = \{s_{ij}\}_{i=1,j=1}^{N,S_N}$, containing N video clips with S_N sentences each, and external knowledge $\mathcal{K} = \{k_m\}_{m=1}^{M}$ with M knowledge sentences, the Knowledge Utilization score is calculated as follows:

$$\text{Know}_{\text{rel}} = \frac{1}{N}(\sum_{i=1}^{N}\frac{1}{S_N}\sum_{j=1}^{S_N}\max_{k\in\mathcal{K}} f_{s_j}^T f_k)$$
$$\text{Know}_{\text{rich}} = \frac{1}{N}\Big|\cup\{k\in\mathcal{K}|\max_{s\in\mathcal{S}} f_s^T f_k > 0.85\}\Big|,$$

where f_s and f_k denote the normalized embeddings of video narration sentences s and external knowledge sentences k.

Word Count Accuracy. Effective narration must align with practical constraints such as video duration and Text-to-Speech (TTS) speed. To capture this requirement, we define a target word count range for each video clip as $[w_i - 3, w_i + 3]$[4], where w_i denotes the ground-truth word count of clip i. During generation, models are guided to produce narrations within this range. We then measure the percentage of outputs that satisfy this constraint, denoted as Hit.

Fluency. Fluent narration requires smooth, grammatically correct, and non-redundant text. To evaluate this property, we follow Guan et al. (2020) and use $Repetition$-4, which measures redundancy in generated outputs. Lower scores indicate higher fluency.

Coherence. Coherent narration requires maintaining logical flow across consecutive clips and avoiding contradictions or abrupt topic shifts. Since no perfect automatic metric currently exists, we employ GPT-4 (Achiam et al., 2023) to rate $Coherence$ on a 1–5 scale, as detailed in Appendix A.3. To reduce inconsistency, we average three independent evaluations following Stureborg et al. (2024). To further ensure reliability, we also assess coherence using additional LLMs and compute correlations between their ratings. The high correlations reported in Appendix A.4 confirm the robustness of our LLM-based evaluation.

4 Methods

Baselines. For the tested MLLMs, we design a straightforward generation pipeline. As detailed in Appendix A.6, the input to each model includes four components: the current video scene, previously generated narrations (for preceding video clips), relevant external knowledge, and word count constraints. The model then generates the narration for the current clip based on these inputs.

VideoNarrationAgent. Our experiments and analysis (§5) reveal three major limitations in current MLLMs for video narration: 1) underuse of previous narrations, leading to incoherent storytelling. 2) repetitive retrieval of similar knowledge across clips, resulting in poor $\text{Know}_{\text{rich}}$. 3) ineffective integration of visual content, narration context, and external knowledge in end-to-end generation.

To address these limitations, we propose **VideoNarrationAgent**, a multi-agent framework composed of three specialized agents (Figure 5): 1) **Narration Context Compression Agent** distills past narrations into concise summaries and key topics. This reduces context length, mitigates redundancy, and improves coherence. 2) **Knowledge Retrieval Agent** retrieves the top five most relevant knowledge sentences for the current visual scene, enhancing informativeness and relevance. 3) **Narration Generation Agent** integrates the current visual clip, compressed narration context, and retrieved knowledge into the final narration, balancing all inputs to produce fluent, coherent, and knowledge-enriched outputs.

Further details of each agent are provided in Appendix A.13, and ablation studies on retrieval strategy and number of retrieved knowledge sentences are reported in Appendix A.12.

5 Experiments and Analysis

We conduct a comprehensive evaluation of eight open-source MLLMs and three closed-source MLLMs on HowToNarrate. Detailed results and analysis are presented in §5.2. We further demonstrate that instruction tuning on our dataset yields significant performance gains. To better understand the task's challenges, we carry out fine-grained analysis in §5.3, and based on these insights, propose a multi-agent framework that achieves further improvements.

5.1 Experimental Settings

Evaluated Models. We evaluate 11 MLLMs: eight open-source and three closed-source. The open-source models include Video-LLaVA (Lin et al., 2023), Video-ChatGPT (Maaz et al., 2023), LongVA (Zhang et al., 2024b), mPLUG-Owl3 (Ye

[4]In practical applications, calculating the word count range based on the required TTS speech style (rate) and video duration allows for flexible control.

		C	M	B-4
Open-Source MLLMs	Video-LLaVA	12.71	8.90	3.88
	Video-ChatGPT	14.67	10.32	4.97
	LongVA	14.46	7.85	2.54
	mPLUG-Owl3	20.33	9.60	4.26
	Chat-UniVi-v1.5	10.39	10.06	2.68
	MiniCPM-V	16.56	11.67	3.51
	Qwen2.5-VL	21.48	9.36	5.29
	InternVL3	**33.98**	**12.26**	**6.58**
Closed-Source MLLMs	Step-1o	52.39	12.42	5.98
	GPT-4o	69.71	14.03	7.04
	Gemini-2.0-Flash	**98.58**	**16.32**	**13.60**
Finetuned MLLM	Qwen2.5-VL	24.79	11.42	5.46
	Qwen2.5-VL*	**144.38**	**17.82**	**17.11**

Table 2: Reference-based Results on HowToNarrate. * indicates finetuned model on HowToNarrate. C: CIDEr, M: METEOR, B-4: BLEU-4.

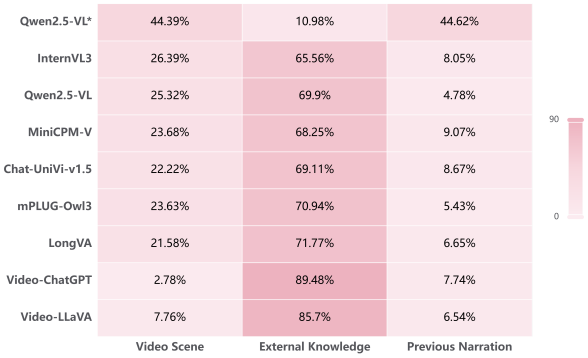


Figure 6: The distribution of attention ratios during narration generation, across three input components: video scene, external knowledge, and previous narrations. Notably, * indicates model finetuned on HowToNarrate.

et al., 2024), Chat-UniVi-v1.5 (Jin et al., 2023), MiniCPM-V (Yao et al., 2024), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025). The closed-source models are Step-1o (Team, 2024), GPT-4o (Hurst et al., 2024), and Gemini-2.0-Flash (Comanici et al., 2025). To validate the quality of HowToNarrate, we additionally fine-tune Qwen2.5-VL. All open-source MLLMs are evaluated using their 7B version, except InternVL3 (8B). We uniformly sample 8 frames per video clip.[5] All models use beam search with num_beams=3 and $temperature$=0.7.

Instruction Tuning. We fine-tune Qwen2.5-VL for 2 epochs with a learning rate of 10^{-4} and a batch size of 32. We use 80% of the data for training and the remaining 20% for testing. LoRA parameters are set to r=64 and α=16. Training takes about 8 hours on 4 NVIDIA A6000 GPUs.

5.2 Main Evaluation Results

Reference-based Results. Table 2 reports the performance of models under traditional similarity metrics on HowToNarrate. Among the open-source models, InternVL3 achieves the strongest results. Closed-source models consistently outperform open-source ones, with Gemini-2.0-Flash ranking the highest overall. Notably, instruction tuning on our dataset further boosts performance: the fine-tuned Qwen2.5-VL surpasses all evaluated models — both open- and closed-source — demonstrating the effectiveness of HowToNarrate.

[5]As detailed in AppendixA.7, larger model sizes and more sampled frames show no significant performance gain, confirming the adequacy of our settings. More evaluated models and perspectives see AppendixA.8

Multi-Dimensional Results. We analyze model performance along five key dimensions: *Visual Relevance, Knowledge Utilization, Word Count Accuracy, Fluency*, and *Coherence*. As shown in Table 3, closed-source models generally out-perform open-source ones. However, they still struggle with controlling narration length and avoiding redundancy, leading to suboptimal fluency and coherence. Fine-tuning on our dataset significantly alleviates these issues, as evidenced by the improved performance of Qwen2.5-VL*.

An interesting observation is that different MLLMs exhibit distinct areas of strength, particularly among open-source models. For coherence and fluency, MiniCPM-V achieves the best performance and also demonstrates the highest knowledge diversity (best $\text{Know}_{\text{rich}}$). In contrast, mPLUG-Owl3 performs strongly in knowledge relevance and word length control, reflecting its stronger capability in handling textual modalities. InternVL3 excels in visual relevance while also maintaining competitive results across other aspects, suggesting its superior ability to integrate visual and textual modalities for narration generation. These findings are further supported by the attention distribution analysis in Figure 6 (see §5.3 for details).

5.3 Fine-Grained Analysis

In synchronized video narration generation, MLLMs are expected to effectively integrate information from three sources: the visual scene, external knowledge, and previous narrations. To better understand model behavior, we analyze the attention distribution using attention rollout (Abnar and Zuidema, 2020), which is computed as

	Visual Relevance			Knowledge Utilization		Word Count Acc.	Fluency	Coherence
	EMScore↑	$\text{EMScore}_{\text{ref}}$↑	Vis_{rel}↑	Know_{rel}↑	$\text{Know}_{\text{rich}}$↑	Hit↑	Repetition-4↓	Coh↑
Video-LLaVA	25.19	55.41	3.02	83.33	43.43	13.04	37.03	2.59
Video-ChatGPT	25.38	55.71	3.00	85.29	53.52	7.70	32.80	2.58
LongVA	25.05	55.45	3.43	87.60	58.88	26.47	37.04	2.90
mPLUG-Owl3	25.39	55.67	3.57	**88.99**	66.66	**28.89**	34.35	2.87
Chat-UniVi-v1.5	25.25	55.46	3.40	85.28	49.27	4.96	28.37	3.10
MiniCPM-V	25.37	55.58	3.78	87.55	**68.56**	6.85	**14.98**	**3.14**
Qwen2.5-VL	25.20	55.49	3.60	83.75	51.72	23.49	38.89	2.59
InternVL3	**25.68**	**56.22**	**3.82**	88.77	67.37	25.73	20.79	2.96
Step-1o	25.94	56.89	4.21	89.05	67.57	52.29	19.26	3.36
GPT-4o	**26.24**	57.20	**4.31**	88.48	67.73	**69.16**	**8.89**	**3.41**
Gemini-2.0-Flash	25.89	**57.30**	4.10	**89.72**	**70.34**	48.04	20.93	3.04
Qwen2.5-VL	**25.50**	55.90	3.62	85.28	52.79	24.76	40.91	2.60
Qwen2.5-VL*	25.48	**57.42**	**4.33**	**87.81**	**70.55**	**84.08**	**6.89**	**3.45**

Table 3: Multi-Dimensional Evaluation (detailed in §3.3) Results. * indicates finetuned model on HowToNarrate.

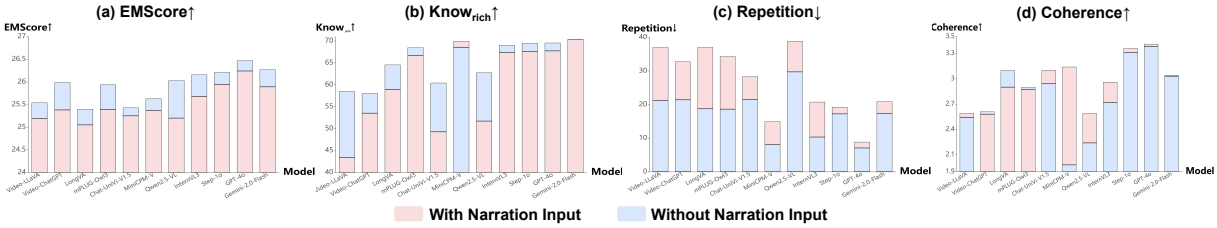


Figure 7: Performance comparison with (pink) and without (blue) inputting previous narrations. The performance gap suggests that current models struggle to utilize prior context, indicating a lack of temporal coherence modeling.

	Agent	CIDEr	METEOR	BLEU-4
Step-1o	-	52.39	12.42	5.98
	✓	**58.86**	**17.32**	**7.98**
GPT-4o	-	69.71	14.03	7.04
	✓	**72.90**	**14.22**	**7.95**
Gemini-2.0-Flash	-	98.58	16.32	13.60
	✓	**125.64**	**18.40**	**14.33**

Table 4: Performance on HowToNarrate demonstrates the clear benefits of our VideoNarrationAgent.

$R = \prod_{l=1}^{L}(A^l + I)$, where A^l is the attention weights of the layer and I is the identity matrix.

Excessive Attention on External Knowledge. As shown in Figure 6, all models pay tremendous attention to external knowledge while largely ignoring the visual scene and previous narration context. Without fine-tuning, over 65% of attention is directed toward the knowledge input. This behavior explains the results in Table 3: although models achieve high knowledge relevance, they often fail to identify knowledge that is both visually relevant and novel relative to previous narrations, leading to high Repetition-4 scores and low $\text{Know}_{\text{rich}}$ performance. Our case study further confirms this: errors frequently occur when consecutive video clips exhibit minimal visual changes, which hinders coherent narration and causes redundant content. To mitigate this, our multi-agent framework incorporates a **Knowledge Retrieval Agent** that selects more visually relevant knowledge and reduces unnecessary information in the final narration. Appendix A.10 provides concrete examples of failure cases and their recovery.

Difficulty in Understanding Narration Context. As shown in Figure 6, all models allocate less than 10% attention to the previous narration, indicating a struggle to leverage past context for coherent storytelling. Instead of building on prior narrations, models may even misinterpret them as external knowledge, leading to repetitive or inconsistent outputs. Our ablation study, comparing model performance with and without access to previous narrations, provides further insights: as Figure 7 shows, removing previous narrations improves Visual Relevance, Repetition, and Knowledge Richness, but at the cost of decreased Coherence.

To address this, our multi-agent framework incorporates a **Narration Context Compression Agent**. This agent distills previous narrations into a concise representation, reducing the difficulty of context understanding while retaining enough information to maintain coherent narration across video clips.

5.4 Proposed Method: VideoNarrationAgent

Based on the insights from §5.3, we propose a multi-agent framework (§4) with three specialized agents: (1) Narration Context Compression Agent,

	Visual Relevance			Knowledge Utilization		Word Count Acc.	Fluency	Coherence
	EMScore↑	$\text{EMScore}_{\text{ref}}$↑	Vis_{rel}↑	Know_{rel}↑	$\text{Know}_{\text{rich}}$↑	Hit↑	Repetition-4↓	Coh↑
Step-1o	25.94	56.89	4.21	89.05	67.57	**52.29**	19.26	3.36
a. w/ Multi-Agents	**26.02**	**57.24**	**4.23**	90.20	**73.04**	51.70	**10.33**	**3.39**
b. w/ KR Agent	25.98	57.18	4.23	**90.25**	72.49	51.77	16.16	3.36
c. w/ NCC Agent	26.02	57.22	4.22	88.86	71.45	51.99	11.06	3.38
GPT-4o	**26.24**	57.20	4.31	88.48	67.73	69.16	8.89	3.41
a. w/ Multi-Agents	26.22	**57.27**	**4.32**	**89.15**	**70.80**	**76.30**	**6.02**	**3.43**
b. w/ KR Agent	26.19	57.22	4.29	88.57	68.43	72.45	6.86	3.42
c. w/ NCC Agent	26.18	57.24	4.31	87.46	67.88	72.39	7.95	3.41
Gemini-2.0-Flash	25.89	57.30	4.10	**89.72**	70.34	48.04	20.93	3.04
a. w/ Multi-Agents	**25.92**	**57.53**	**4.13**	88.46	**71.36**	74.62	**5.76**	**3.43**
b. w/ KR Agent	25.86	57.47	4.12	87.53	70.66	**74.71**	7.59	3.34
c. w/ NCC Agent	25.83	57.35	4.11	87.46	71.17	70.02	5.86	3.35

Table 5: Ablation study of each agent in our VideoNarrationAgent. The best result on each aspect is **bolded** and the second best is underlined.

	Vis_{rel}	Coh	Att
Qwen2.5-VL	3.42	2.81	2.78
Qwen2.5-VL (w/ Finetuned)	**4.05**	**3.64**	**3.89**
Gemini-2.0-Flash	3.86	3.32	3.24
Gemini-2.0-Flash (w/ Multi-Agents)	3.90	3.59	3.29

Table 6: Human evaluation results on three aspects.

(2) Knowledge Retrieval Agent, and (3) Narration Generation Agent, targeting coherence, repetition, and knowledge utilization issues in current MLLMs. Table 4 shows that our method substantially improves overall performance. Ablation studies in Table 5 confirm each agent's contribution: the Knowledge Retrieval Agent boosts $\text{Know}_{\text{rich}}$, while the Narration Context Compression Agent enhances Fluency and Coherence.

5.5 Human Evaluation

We conduct human evaluations across three metrics: (1) *Visual Relevance* measures the degree of alignment between the generated narrations and the corresponding video shots. (2) *Coherence* measures the internal logical consistency of the generated narrations. (3) *Attractiveness* measures the extent to which the narrations improve the users' viewing experience. All metrics are rated on 1–5 scale. The results are shown in Table 6 (see §A.15 for details). It demonstrates that both our finetuning and multi-agents methods are effective.

6 Related Works

Video Narration. Traditional video captioning generates bland summaries, unsuitable as engaging voiceovers. Recent works aim to create more compelling narratives (Li et al., 2019). Shot2Story20K (Han et al., 2023) offers single-shot narrations but lacks logical connections. Bhattacharya et al. (Bhattacharya et al., 2023) annotated advertising videos with behavioral and emotional dimensions, but their dataset is limited in scale and domain. Yang et al. (Yang et al., 2024) introduced Synchronized Video Narration with fine-grained annotations for advertising, yet it lacks generalizability. We contribute a general-domain dataset with longer videos, unstructured knowledge, and creative narrations, introducing new challenges.

MLLM-based Video-to-Text Generation. MLLMs are increasingly applied to video-to-text tasks but still face challenges in fine-grained understanding. Multi-stage frameworks tackle more complex tasks requiring knowledge incorporation and coherence. ShareGPT4Video (Chen et al., 2024a) generates differential captions, then summarizes. MM-Narrator (Zhang et al., 2024a) detects characters and captions videos before narrating movies. However, these methods lack input filtering and often attend to irrelevant information. Our approach addresses this via knowledge retrieval and context compression, effectively focusing on relevance.

7 Conclusion

This paper presents a comprehensive study on synchronized video narration. We introduce HowToNarrate, the first general-domain synchronized video narration dataset. Our evaluation of current MLLMs across multiple dimensions reveals three major bottlenecks: (1) repetitive knowledge retrieval, (2) insufficient modeling of contextual narrations, and (3) challenges in integrating multiple information sources. To address these issues, we propose VideoNarrationAgent, which integrates three agents to enhance video narration generation: Narration Context Compression Agent, Knowledge Retrieval Agent, and Narration Generation Agent.

Experiments show that our method significantly improves video narration performance.

Limitations

The provision of external knowledge is crucial to enhance the information richness of generated video narration. In this study, we mainly focus on generating synchronized video narration based on already-given external knowledge, which is a step towards this goal. However, in real life, how to obtain such knowledge is also crucial. To solve this problem, we can further improve our system through RAG (Retrieval-Augmented Generation). We make a simple attempt in Appendix A.8.2.

Ethics Statement

We propose HowToNarrate, a new benchmark, to support the exploration of Synchronized Video Narration. There are two potential ethical issues with our work, which are related to the data source and crowdsourcing services. We state each of them as follows:

Data source. The collected videos are all from YouTube and crawled according to the service contract of the website. Considering copyright issues, we only publish a list of URLs of the videos. In addition to providing URL lists, we will consider sending videos with missing links to applicants for non-commercial purposes via email, after rigorously reviewing their intended use. It greatly ensures the reproducibility of HowToNarrate. Our data source does not contain any names, personally identifiable information, or offensive content.

Crowdsourcing services. After optimizing the data using the powerful DeepSeek-V2.5, we only need to further check the annotations. We have hired 5 workers (college students, aged 22–29) to review and correct any remaining errors. Each video takes about 5 minutes to complete. During human evaluation, we have hired 10 annotators (college students, aged 21–26) included 6 women and 4 men. Each annotator evaluate video narrations and give their rating from 1 to 5. Each video takes about 2 minutes to complete. All of the above-mentioned workers received reasonable compensation in line with local labor standards.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62576347).

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. *arXiv preprint arXiv:2305.09758*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. 2012. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE transactions on visualization and computer graphics*, 18(12):2649–2658.

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

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2024b. Longvila: Scaling long-context visual language models for long videos. *Preprint*, arXiv:2408.10188.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and

next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 968–974.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. 2023. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17929–17938.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Yidan Sun, Jianfei Yu, and Boyang Li. 2024. Multilingual synopses of movie narratives: A dataset for story understanding. *arXiv e-prints*, pages arXiv–2406.

Jieyue Xingchen Team. 2024. Step-1o: A multimodal language model. *Accessed: 2024-10-01*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. *arXiv preprint arXiv:2405.14040*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024a. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13657.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Details of Dataset Construction

We have hired 5 workers (college students, aged 22–29) to review and correct any annotation errors. Each video takes about 5 minutes to complete.

A.1.1 ASR refinement

ASR refinement proceeds in two steps:

- **DeepSeek-V2.5 performs preliminary refinement.** It corrects homophones, incoherence, channel information, and other errors. At this stage, the auto-ASR results contain an initial error rate of 15.3%.
- **Human refinement.** After DeepSeek-V2.5 corrects typos and major errors in key information, the error rate decreases to 6.7%, and human annotators refine any remaining errors.

The prompt for ASR refinement is shown in Figure 10.

A.1.2 External Knowledge Construction

Our constructed knowledge consists of two components:

- **Knowledge from the ground truth human-written narrations.** It ensures our provided knowledge covers the most suitable knowledge for the video. This part contains minor hallucinations, as verified by our annotators.
- **Extended knowledge generated by the LLM.** It provides distractor or possibly usable knowledge for narration generation. For instance, in a video showing camel humps, the GT knowledge is "saving fat," while extended knowledge like "enhance heat dissipation" can also be used for this clip, whereas extended "courtship" knowledge serves as a distractor. All this extended knowledge is provided to challenge the model's ability to retrieve suitable knowledge from the extended set, not just easily identify the correlated "saving fat"; This part may contain potential prior knowledge or hallucination issues. We performed human verification on all results, with a pass rate of 87.3%. However, we clarify that slight hallucinations do not affect our evaluation of knowledge retrieval ability, as their correlation with the video (whether "suitable") is unquestionable.

Our goal of expanding external knowledge is to supplement the knowledge points not covered in the ground truth (GT) narration by incorporating the world knowledge from the LLM. It covers all the knowledge in our narration annotations while preventing the model from mechanically using all the knowledge we provide in order to achieve higher metrics. In other words, our ultimate goal is for models to selectively integrate certain information from knowledge based on the content of the video, which places greater demands on the model's multimodal integration capabilities.

The extraction of external knowledge is shown in Figure 11, and the expansion of external knowledge is shown in Figure 12.

A.2 More Data Cases

The data cases in the domains of Handcraft, Travel, and Engneering Science are shown in Figure 14-16.

A.3 Prompts for LLM-based evaluation

The prompt for visual relevance evaluation is shown in Figure 9, and the prompt for coherence evaluation is shown in Figure 8.

A.4 Reliability of LLM-based evaluation

To avoid potential bias of Gemini-2.5-pro-based visual relevance evaluation due to model preference, we ensure that Gemini-2.5-pro, which we use for evaluation, is not included among the tested generation models. To further assess potential bias, we conduct evaluation experiments using multiple MLLMs (Gemini-2.5-pro and GPT-4o). Table 7 presents the Pearson correlation of evaluation results between these different MLLMs, where the correlation of coherence scores across models is high, indicating the reliability of our evaluation method.

Similarly, to vertify GPT4-based coherence evaluation, we conducted evaluation experiments using multiple LLMs (GPT-4, GPT-5, Gemini-2.5-pro, and Deepseek-chat). Table 8 presents the Pearson correlation of evaluation results between these

Model Pairs	InternVL3's generations	Qwen2.5-VL's generations
GPT-4o & Gemini-2.5-pro	0.65	0.64

Table 7: The Pearson correlation of visual relevance evaluation results between different MLLMs.

Model Pairs	InternVL3's generations	Qwen2.5-VL's generations
GPT-4 & GPT-5	0.70	0.75
GPT-4 & Deepseek-chat	0.65	0.77
GPT-4 & Gemini-2.5-pro	0.63	0.66
GPT-5 & Deepseek-chat	0.65	0.62
GPT-5 & Gemini-2.5-pro	0.65	0.62
Deepseek-chat & Gemini-2.5-pro	0.65	0.66

Table 8: The Pearson correlation of coherence evaluation results between different LLMs.

different LLMs. Most of the results above 0.65 indicate the reliability of our evaluation method.

A.5 Discussion of the Visual Contribution

To quantify the contribution of visual scenes to narration, we performed an ablation using only text-based input for Gemini-2.5-Flash and evaluated the Visual Relevance of the outputs. As shown in Table 9, performance dropped significantly across all metrics. These results clearly demonstrate the crucial role of visual input in generating high-quality narrations.

A.6 Prompt for Our Tested LLMs

The prompt for our tested MLLMs is shown in Figure 17.

A.7 Ablation Studies on Model Parameters and the Number of Sampled Frames

We conducted additional experiments with Chat-UniVi-v1.5 and InternVL3 using larger parameter models. The results in Table 10 show that doubling the parameters don't clearly improve performance on this task. This suggests that current MLLMs' training data lacks tasks requiring a comprehensive understanding of textual knowledge, visual elements, and textual context in continuing generation. Specifically, these models struggle to distinguish between textual knowledge and textual context. And such abilities cannot emerge through parameter scaling alone.

According to statistics, the length of each video clip is not long (with an average length of 12.3 s). Thus, 8 frames could be sufficient. We conduct further ablation studies on the number of frames using LongVA and mPLUG-Owl3 in Table 11, and find that increasing the number of sampled frames don't improve the results, verifying that our current sampling number is adequate.

	EMScore	$\text{EMScore}_{\text{ref}}$	Vis_{rel}
w/o visual inputs	21.64	52.31	2.51
w/ visual inputs	**25.89**	**57.30**	**4.10**

Table 9: Visual relevance evaluation results of Gemini-2.5-pro w/ and w/o visual input. The best performance for each metric is **bolded**.

A.8 The Results of More Evaluated Models and Perspectives

A.8.1 More Evaluated Models

We added the evaluation experiments for LongVILA (Chen et al., 2024b) and LongVU (Shen et al., 2024) in Table 13 and Table 12. Compared to other open-source models, LongVILA performs strongly in knowledge relevance and word length control, reflecting its stronger capability in handling visual modalities. However, its overall performance still does not surpass InternVL3. In comparison, LongVU's abilities in areas other than visual relevance are somewhat mediocre.

A.8.2 Evaluation of RAG for Scaling up Knowledge.

Although the current knowledge base already challenges models to find suitable knowledge, further intensifying this challenge will provide more insights. Here, we use RAG to extend the current knowledge base to at least three times its original size. We select three strong models that performed well on the original knowledge base, and the experimental results are shown in Table 15. InternVL3 and GPT-4o both show performance degradation, while Gemini-2.0-Flash demonstrates superior long-context processing capabilities. Additionally, beyond the difficulty of knowledge retrieval, we also emphasized the limitations of underusing narration history and the difficulty in integrating multimodal information. These three challenges collectively constitute this complex task.

A.9 The Generalization of the Finetuned Model

Since our dataset is the first English-language dataset for synchronized video narration, there are no suitable benchmarks to test generalization. As an alternative, we evaluate domain generalization. Specifically, we remove Documentary and Handcraft domain data from the training set and infer-

	Parameter	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
InternVL3	8B	33.98	12.26	28.27	14.45	9.23	6.58
	14B	43.93	13.89	29.32	14.59	8.87	6.02
Chat-UniVi-v1.5	7B	10.39	10.06	21.80	8.43	4.36	2.68
	14B	13.87	11.32	23.29	9.75	5.24	3.30

Table 10: Performance comparison across MLLMs on different model parameters. Larger model parameters don't lead to improved performance.

	Sampled Frames	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
LongVA	8 frames	14.46	7.85	20.06	7.53	4.00	2.54
	16 frames	13.27	7.61	19.70	7.32	3.89	2.54
mPLUG-Owl3	8 frames	20.33	9.60	24.60	10.87	6.34	4.26
	16 frames	20.90	9.64	24.57	11.07	6.47	4.40

Table 11: Performance comparison across MLLMs on different numbers of sampled frames. Larger sampled frames don't lead to improved performance.

ence on these two domains using finetuned and non-finetuned models. As shown in Table 16, the finetuned model demonstrates strong generalization performance on unseen domains, significantly outperforming baseline models.

A.10 Failure and Recovery Cases

Our case studies show that errors often arise when adjacent video clips lack distinct visual changes, leading to redundant content. As explained in our attention analysis, baseline models pay limited attention to prior narrations, preventing coherent continuation and causing repetition. Our proposed method alleviates this issue. The concrete failure and recovery cases are shown in Figure 13.

A.11 Discussion of Attention Shift after Finetuning

We suggest that for this task, the model primarily relies on visual input and previous narration to generate current narration, rather than needing to process large amounts of redundant external knowledge. Through finetuning, the model effectively learned the appropriate attention allocation patterns required for this task, and achieved significantly better overall performance. This attention-based interpretability analysis and the demonstrated effectiveness of our framework provide valuable insights for future research.

A.12 Ablation Studies on Knowledge Retrieval Agent

Before finalizing our method, we conducted ablations on semantic-similarity–based retrieval. Results in Table 17 show that such retrieval underperforms compared to leveraging MLLMs directly, since similarity matching captures only surface meaning, whereas MLLMs enable deeper logical reasoning. As shown in Table 18, we also tested retrieving 3, 5, and 7 knowledge sentences, finding that too few or too many degraded performance. Retrieving 5 sentences yield the most balanced results.

A.13 Prompts for VideoNarrationAgent

The prompt for Narration Context Compression Agent is shown in Figure 18, the prompt for Knowledge Retrieval Agent is shown in Figure 19, and the prompt for Narration Generation Agent is shown in Figure 20.

A.14 Discussion of Inconsistent Improvements across Different Models with Multi-Agent Framework

Different models vary in capabilities across tasks, leading to uneven improvements. For instance, unlike other models that show clear improvement in word length control, Step-1o does not show significant improvement due to its limited word length control ability, and cannot be enhanced by breaking down the original complex task. Nevertheless, our ablation studies in Table 5 clearly demonstrate that the Knowledge Retrieval Agent enhances knowledge utilization across all models, and the Narration Context Compress Agent significantly improves both fluency and coherence.

A.15 Details of Human Evaluation

All the annotators are graduate students with master's degree or higher. The team of 10 annotators included 6 women and 4 men, ranging in age from 21 to 26. Each annotator evaluate 300 narrations

	Visual Relevance			Knowledge Utilization		Word Count Acc.	Fluency	Coherence
	EMScore↑	$\text{EMScore}_{\text{ref}}$↑	Vis_{rel}↑	Know_{rel}↑	$\text{Know}_{\text{rich}}$↑	Hit↑	Repetition-4↓	Coh↑
LongVILA	25.49	56.00	3.67	86.16	55.72	27.05	34.22	2.86
LongVU	25.86	56.13	3.77	84.38	46.94	20.86	36.69	2.87

Table 12: Multi-Dimensional Evaluation Results.

Role:
You act as a text evaluation specialist who assesses video voice-over transcripts for coherence and fluency. You will be provided with a piece of **Voice-Over Text**. Your task is to examine the transcript, focusing on how logical, consistent, and fluent it is when read or spoken aloud. Then, assign a **Single Numeric Score** from 1 to 5, where 1 indicates a highly disjointed and confusing transcript, and 5 indicates a very coherent, smoothly flowing text.

Instructions:
- Carefully evaluate the text for clarity, grammatical consistency, and overall logical progression.
- Provide your final rating in JSON format with the key "coherence_score" as shown below:
  {"coherence_score": <score (1-5)>}
- Do not include any additional commentary or text beyond the JSON output.

Here is the Voice-Over Text:
Please give your evaluation:

Figure 8: The prompt for the coherence evaluation.

	CIDEr	METEOR	BLEU-4
LongVILA	24.77	9.18	5.09
LongVU	14.59	7.05	2.52

Table 13: Reference-based Results on HowToNarrate.

Model	Vis_{rel}	Coh
Qwen2.5-VL	0.65	0.64
Qwen2.5-VL (w/ Finetuned)	0.59	0.63
Gemini-2.0-Flash	0.62	0.64
Gemini-2.0-Flash (w/ Multi-Agents)	0.60	0.60

Table 14: The Pearson correlation between human-evaluated scores and the corresponding LLM-based evaluation scores.

generated by four models based on 75 clips from 10 randomly selected videos. They read the detailed annotation documentation and have a thorough understanding of the synchronized video narration task. The Pearson Correlation score of their evaluation results is 0.63. And the Pearson correlation between human-evaluated scores and the corresponding LLM-based evaluation scores is shown in Table 14. Correlation scores mostly above 0.65 indicate strong reliability of our LLM-based evaluation (a Pearson score greater than 0 indicates positive correlation, and greater than 0.5 indicates strong positive correlation).

	Knowledge Base Size	CIDER	METEOR	BLEU-4	Know_{rel}	$\text{Know}_{\text{rich}}$
InternVL3	1x	33.98	12.26	6.58	88.77	67.37
InternVL3	3x	21.74	10.49	4.05	87.48	66.12
GPT-4o	1x	69.71	14.03	7.04	88.48	69.25
GPT-4o	3x	60.37	12.52	5.88	86.90	67.73
Gemini-2.0-Flash	1x	98.58	16.32	13.60	89.72	71.70
Gemini-2.0-Flash	3x	96.76	16.25	13.84	88.14	70.34

Table 15: Evaluation of RAG for Scaling up Knowledge.

	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Qwen-2.5-VL	19.54	9.18	23.44	10.23	5.77	3.72
Qwen-2.5-VL*	**69.46**	**13.64**	**31.43**	**18.29**	**12.91**	**9.87**

Table 16: This table shows the generalization of the finetuned model on HowtoNarrate, and * indicates that the model has been finetuned. The best performance for each metric is **bolded**.

	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
w/ Semantic-Similarity–Based Retrieval	112.08	16.72	36.88	22.76	16.27	12.48
w/ MLLM–based Retrieval	**125.64**	**18.40**	**39.58**	**25.47**	**18.57**	**14.33**

Table 17: This table shows the performance comparison between semantic-similarity–based retrieval and MLLM-based retrieval in Knowledge Retrieval Agent on Gemini-2.0-Flash. The best performance for each metric is **bolded**.

	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
w/ 3 retrieved sentences	124.10	18.21	39.02	25.04	18.25	13.16
w/ 5 retrieved sentences	**125.64**	**18.40**	**39.58**	**25.47**	**18.57**	**14.33**
w/ 7 retrieved sentences	119.80	17.22	37.26	24.13	16.35	12.27

Table 18: This table shows the performance comparison between different numbers of retrieved knowledge sentences in Knowledge Retrieval Agent on Gemini-2.0-Flash. The best performance for each metric is **bolded**.

Role:
You are an expert in multimodal content evaluation, specializing in video narration evaluation. Your task is to evaluate the effectiveness of a given **Video Narration** with respect to its corresponding **Video Scene**, based on the following evaluation criteria.

Evaluation Criteria:
- Content Relevance -: Is the narrative content logically or indirectly relevant to the video scene?

Rating:
Assign a rating between 1 and 5 to each of the above evaluation criteria, where:
- 1 -: Poor - the narration is largely irrelevant or contradictory to the visuals.
- 2 -: Fair - the narraton is partially relevant but has major flaws.
- 3 -: Good - the narration is generally appropriate but has some minor issues.
- 4 -: Very Good - the narration fits the visuals very well with few negligible issues.
- 5 -: Excellent - the narration perfectly complements and enhances the visuals in every way.

Response Format:
Please provide your evaluation strictly in the following JSON format:

```
{
  "Content Relevance": <score (1-5)>
}
```

- Do not add any additional text, explanations, or comments outside the specified format.
- Make sure the score reflects an overall evaluation based solely on the video and narrative provided, and do not use any external knowledge or assumptions beyond the given material.

Here is the Video Scene:
Here is the corresponding Video Narration:
Please give your evaluation:

Figure 9: The prompt for visual relevance evaluation.

You act as a professional editor specializing in video transcription cleaning and correction. Given a JSON file with the following structure:

{"segments": [{"start": start_time, "end": end_time, "text": narrative_text}, ...]}

Please process the data according to the following requirements:

1. For each object in the segments array, focus on the "text" field (the speech recognition result of the video narration).

2. Remove any content related to youtuber name, channel promotion, subscription reminders, welcome messages, thank-you notes or farewells from the "text" field. This includes phrases like "Welcome back to my channel," "Don't forget to subscribe," "Please like and follow," "See you next time," etc.
- If the entire "text" field consists solely of such content, delete the entire object from the "segments" array.

3. Correct any disfluent filler words, ellipses (such as "um," "ah," "you know") and common errors from automatic speech recognition within the "text" field.

4. Appropriately optimize the sentences while preserving their original meaning as much as possible, ensuring they are coherent, and naturally connected to the preceding text.

5. Ensure that the revised sentences are linguistically fluent, logically coherent, have accurate punctuation, and that the overall narration remains cohesive.

6. Do not alter the values of the "start" and "end" fields. After processing, output the resulting JSON data, maintaining the original structure and correct formatting.

7. When processing, only focus on the content of the current segment's "text" field. Do not reference or merge content from the "text" fields of previous or subsequent segments.

Please output only the resulting JSON data, without any additional explanations.

Here is the JSON data to be processed:

Figure 10: The prompt for the ASR refinement.

You are a data extraction expert who specializes in video transcription. Given a JSON file with the following structure:

[{"id": "<id>", "video_path": "<video_path>", "frames": [{"start": <start_time>, "end": <end_time>, "text": "<transcribed_text>"}, ...]}]

Your tasks are:

1. Extract key information:
- Analyze and extract professional information and external knowledge points present in all `text` fields within the `frames` array of the JSON file.
- Focus on informative and useful content. Exclude colloquial and informal expressions.

2. Maintain accuracy and fidelity:
- Ensure that the extracted information is faithful to the original.
- Accurately reflect the content without adding any unnecessary interpretations or distortions.

3. Present knowledge points in a specific format:
- First knowledge point: Determine the overall topic of the video based on the entire content. Titled "Topic", presented as the first knowledge point.
- Other knowledge points: Identify and categorize different knowledge points. Make sure each knowledge point represents different information or concepts, and make sure each knowledge point has a clear definition and is separate from other knowledge points. Each knowledge point should be independent, self-contained, and understandable without additional context.
- Structure: Each knowledge point starts with a clear and descriptive title related to the topic. The title is followed by a detailed description or explanation. Use clear and fluent language for easy understanding.

4. Output format:
- Provide the final result in the following JSON format:
[{"knowledge": "<extracted_knowledge_points>"}]
where the `knowledge` field contains all the extracted knowledge points, presented with title and content. The title and content of each knowledge point are separated by ":", and the newline break `\n` is used to separate different knowledge points. The format is as follows:
```
Title of knowledge point 1: detailed description or explanation of knowledge point 1.
Title of knowledge point 2: detailed description or explanation of knowledge point 2.
Title of knowledge point 3: detailed description or explanation of knowledge point 3.
```

Here is the JSON data to be processed:

Figure 11: The prompt for the extraction of external knowledge.

Role:
You are a knowledge architect.

Task:
Your task is to expand and supplement the existing knowledge outline I provided. The goal is to make this outline more comprehensive, in-depth and structured.

Existing knowledge outline:
{Initial External Knowledge}

Expansion requirements:
1. Maintain consistency: All new content must be highly consistent with the core ideas and logic of the original text, and no contradictions or conflicts should occur.
2. Deepen existing knowledge points: For existing knowledge points in the outline, if possible, please add more detailed explanations, specific examples, relevant background information or subdivided sub-knowledge points.
3. Supplement new knowledge points: Under each category, add 2-3 new knowledge points that are closely related to the theme of the category but not mentioned in the original text.
4. Maintain structure: Strictly abide by and continue the hierarchy and format of knowledge point provided by me.

Output format:
Please modify and expand directly based on the original outline, and return a complete new outline using the format I specify.

Figure 12: The prompt for the expansion of external knowledge.

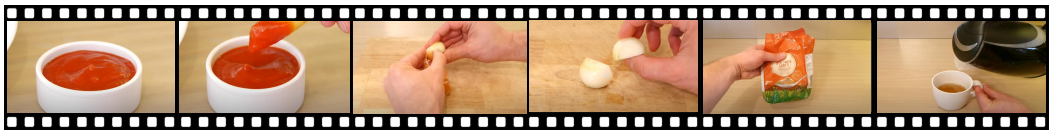
00:00 - 00:09 (9s)	00:09 - 00:16 (7s)	00:16 - 00:28 (8s)
(a) Failure Case		
Males often erupt in prolonged and fierce conflicts to contend for mating rights.	Facing the harsh desert environment, male Nubian ibexes often get drawn into lengthy duels to win the affection of females.	Male Nubian ibex fiercely compete, enduring battles for female attention in harsh conditions.
(b) Recovery Case		
For a male Nubian ibex, courtship is highly competitive.	On these barren slopes in eastern Israel, life is tough at the best of times. As the temperatures soar, it's about to get tougher.	The females need to give birth in spring, so they must mate soon.

00:00 - 00:05 (5s)	00:05 - 00:11 (6s)	00:11 - 00:15 (4s)
(a) Failure Case		
Pour the coffee into a mug, then gently spoon the creamy egg mixture over the coffee in the mug.	Pour boiling water over the coffee grounds in the filter.	Boiling water water drips into a clear glass mug.
(b) Recovery Case		
Now, add 2 tablespoons of boiling water and check the filter. The coffee should drip through slowly. If it's coming out too fast, tighten the filter.	Add the remaining boiling water and allow the coffee to drip through slowly.	For a shortcut, make espresso instead of filtered coffee.

Figure 13: The failure and recovery case.

External Knowledge

Task: How to make homemade tomato ketchup.
Description: This tutorial teaches you how to create your own tomato ketchup from scratch, using simple ingredients and basic kitchen tools.
Steps:
1. Peel a onion and using half of it.
2. Prepare a mixture of brown sugar and hot water.
3. Once the sugar water has cooled, add the half onion to the cup and blend it using a hand blender until smooth.
4. In a separate bowl, mix tomato paste with a quarter teaspoon each of ground cloves, ground cinnamon, and salt, as well as half a teaspoon of mustard.
5. Add three tablespoons of white vinegar and the blended onion-sugar water to the tomato paste mixture. Whisk everything together thoroughly.
6. Adjust the consistency by adding less water if you prefer a thicker ketchup.



00:00 - 00:03 (3s)	00:03 - 00:08 (5s)	00:08 - 00:14 (6s)
Today, I'm going to show you how to make your own tomato ketchup.	Start by taking a small onion and peel it. We're only going to use about half of it.	Next, take some brown sugar and put one heaped tablespoon into a cup. Then fill the cup halfway with hot water so the sugar dissolves.

Figure 14: The data case of how to make homemade tomato ketchup.

External Knowledge

Topic: Tongariro National Park
Natural Features and Landscapes:
1. Tongariro National Park is New Zealand's oldest national park and a UNESCO World Heritage Site, established in 1887.
2. The park features a dramatic volcanic landscape with three active volcanoes: Tongariro, Ngauruhoe, and Ruapehu.
3. The Emerald Lakes are iconic features of the park, known for their vibrant turquoise color due to minerals dissolved in the water, a result of volcanic activity.
Hiking and Outdoor Activities:
1. The Tongariro Alpine Crossing is a 19.4 km trek often considered one of the best day hikes in the world. It passes through diverse landscapes including old lava flows, steam vents, and crater lakes.
2. The park offers numerous shorter walks and viewpoints for less strenuous adventures, such as Taranaki Falls.

00:00 - 00:13 (13s)	00:13 - 00:19 (6s)	00:19 - 00:29 (10s)
The dramatic volcanic landscape of Tongariro National Park, with its three active volcanoes, Tongariro, Ngauruhoe, and Ruapehu, creates a scenery that's truly out of this world.	At the heart of the park lies the Tongariro Alpine Crossing, often hailed as one of the best day hikes in the world.	As you make your way along the trail, you'll be treated to breathtaking views of Emerald Lakes. These stunning turquoise pools nestled in the volcanic terrain, a vivid reminder of the area's volcanic activity.

Figure 15: The data case of Tongariro National Park.

External Knowledge

Topic: The process of weaving and fabric production in the fashion industry.

Weaving Techniques and Processes:

1. The art of weaving has been around since the Stone Age, with traditional weavers using techniques and designs that reflect their cultural origins.
2. Shuttle looms and rapier looms interweave threads under tension, with rapier looms being significantly faster than hand-operated looms.
3. The warp, a set of threads for woven cloth, is prepared on a warping machine, with rolls that can stretch several miles long.
4. Completed fabrics undergo visual quality control, with workers repairing any broken threads by hand.
5. Fabrics go through a two-hour cleaning process, steaming, and a finishing process.

Fiber and Yarn Production:

1. Cashmere goat hair undergoes processing at wool mills, where short fibers called flocks are carded to open up fibers and mix colors.
2. The core function of the carding is to remove impurities and process them into a uniform thin fiber web.
3. Spinning machines separate fibers into threads and wind them onto bobbins, with operators loading yarn into dyeing machines.
5. Uncarded locks of hair, called wool tops, are used to make worsted yarn, with blending machines merging multiple threads into one large strip.
6. The coning machine winds single threads around cones

00:00 - 00:05 (5s)	00:05 - 00:11 (6s)	00:11 - 00:15 (4s)
This is what cashmere goat hair looks like when it arrives at the wool mill.	Short fibers called flocks go through a carding machine, which opens up the fibers and mixes colors together.	The carding process removes impurities and reduces the fibers to a flat layer.

Figure 16: The data case of the process of weaving and fabric production in the fashion industry.

Instructions:
You are a video narration editor who is good at writing engaging narration copy based on video content and can accurately understand external knowledge and video information.
I will provide you with information about the video shot and the corresponding external knowledge. Your task is to combine external knowledge to write a narration copy that meets the word count requirements for the current video shot to support and enhance the visual effect. You need to ensure that the copy is closely related to the visual scene and is coherent with the previous copy.
Please note that when writing the narration copy, the provided external knowledge and the scene information of the shot are mainly used.

Input:
Previous narration:
{previous_narration}
External knowledge:
{external_knowledge}
Video shot:
{video_shot}
Word count requirement:
{word_count_requirement}

Response:

Figure 17: The prompt for our tested MLLMs.

Role:
You are a professional narrative logic analyst responsible for maintaining the contextual coherence of the video narration.

Task:
Based on the Historical Narration Record provided to you, analyze and summarize the current narrative state. Your output will help the narration generator avoid repetition and create logical transitions.

Input:
You will receive a JSON object with the following fields:

```
{
  "historical_narrations": [
    "Narration 1: This is the beginning of our journey...",
    "Narration 2: We first saw the majestic mountains...",
    "..."
  ]
}
```

Output format (JSON):
Please strictly follow the following JSON structure for output:

```
{
  "summary": "Summarize the core narrative content so far in one sentence.",
  "key_entities": "List the core entities (people, things, objects, places) that have been mentioned.",
  "last_narration_focus": "Indicate what the core focus of the last narration is."
}
```

Rules and constraints:
1. Focus on history: All your analysis must be based on the provided Historical Narration Record.
2. High-level summary: The output should be concise and grasp the main points.
3. Logical Analysis: Focus on analyzing the logical relationship between the narrations rather than simply retelling them.

Here is my Historical Narration Record:

Figure 18: The prompt for Narration Context Compression Agent.

Role and Objectives:
You are an advanced knowledge retrieval and content analysis agent. Your core task is to serve as an intelligent narration assistant for video content. Based on the provided video scenes, you can filter out the most valuable supplementary information for the audience from the specified knowledge base.

External Knowledge Base:
{External Knowledge Base}

Core Task Instructions:
Please strictly follow the steps below to retrieve the first five knowledge points from the External Knowledge Base above that best match the current Continuous Video Frames content.

Execution Steps:
1. Visual Analysis:
- Please first analyze the 8 continuous video frames in detail.
- Identify the key objects, environments, and the core actions or state changes that are taking place.
- Comprehensively judge the core scenes and events shown in this set of images.

2. Knowledge association and preliminary screening:
- Match the visual information you observed in step 1 (especially core actions and key objects) with each knowledge point in External Knowledge Base.
- Find a list of candidate knowledge points that are highly relevant to the content of the picture.

3. Sorting and selection:
- For the remaining knowledge points, sort them in descending order according to the following two dimensions:
- Dimension 1: Direct relevance to the picture: Can the knowledge point directly explain or deepen the understanding of the core action or state change in the picture? The more direct the relevance, the higher the ranking.
- Dimension 2: Importance and value: How important is this knowledge point to the audience's comprehensive understanding of the current scene? Is it a key rule, core skill or important precaution? The more important, the higher the ranking.
- Combining the above two dimensions, select the five knowledge points with the highest ranking.

Output format requirements:
Please strictly output the final results in the following JSON format, and do not include any additional explanations, titles or captions.

```json
{
  "retrieved_knowledge": [
    "The first most relevant knowledge point",
    "The second most relevant knowledge point",
    "The third most relevant knowledge point",
    "The fourth most relevant knowledge point",
    "The fifth most relevant knowledge point"
  ]
}
```

Figure 19: The prompt for Knowledge Retrieval Agent.

Role:
You are a top video narration writer who is good at integrating scattered information into smooth, vivid and fascinating narrations. Your language style is elegant, precise and vivid.

Task:
Based on the Video Scenes (8 video keyframes arranged in continuous order) and Creation Brief I provided, create a video narration that meets the Word Count Requirements.

Input: Creation Brief (JSON)
You will receive a JSON object containing all of the following information:

```
{
   "context_summary": { /* Full JSON output from the Context Memory Agent, possibly None */},
   "external_knowledge": { /* Full JSON output from the Knowledge Retrieval Agent */},
   "word_requirement": "From xx to xx words"
}
```

Output Format (JSON):
Please output the following JSON structure closely:

```
{
   "narration_text": "The final narration text you created. This should be a complete, ready-to-use sentence."
}
```

Creation Notes and Thoughts:
Before generating your final narration, think through the following steps in your mind (no need to output your thought process):

1. Analyze the visuals: First, look at Video Scenes. This forms the background of the narration.
2. Connect context: Review `context_summary`. If so, how should I follow up from the previous paragraph? Continue the topic or introduce a new angle? Avoid repeating what is mentioned in `context_summary` unless it is for emphasis.
3. Integrate knowledge: Review `external_knowledge`. Based on your understanding of the visual effect (step 1) and the narrative context (step 2), find the most relevant facts from the provided external_knowledge. Find the most natural and smooth way to integrate this knowledge into the narrative. The goal is to enrich the visual story rather than just state the facts.
4. Integrate within specified word count: Synthesize the insights from the previous steps into a single, coherent narrative within the specified word count, according to the requirements of `word_requirement`.

Rules and constraints:

1. Strictly follow the word count: Strictly follow the specified `word_requirement`.
2. No fiction: Do not invent information that is not provided in the Creation Brief.
3. Strict format: only one valid JSON must be returned.

Here is my Creation Brief:

Figure 20: The prompt for Narration Generation Agent.