ShotFinder: Imagination-Driven Open-Domain Video Shot Retrieval via Web Search

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have made rapid progress in information retrieval, yet existing research has mainly focused on text or static multimodal settings. Open-domain video shot retrieval, which involves richer temporal structure and more complex semantics, still lacks systematic benchmarks and analysis. To fill this gap, we introduce ShotFinder, a benchmark that formalizes editing requirements as keyframe-oriented shot descriptions and introduces five types of controllable single-factor constraints: Temporal order, Color, Visual style, Audio, and Resolution. We curate 1,210 high-quality samples from YouTube across 20 thematic categories, using large models for generation with human verification. Based on the benchmark, we propose ShotFinder, a text-driven three-stage retrieval and localization pipeline: (1) query expansion via video imagination, (2) candidate video retrieval with a search engine, and (3) description-guided temporal localization. Experiments on multiple closed-source and open-source models reveal a significant gap to human performance, with clear imbalance across constraints: temporal localization is relatively tractable, while color and visual style remain major challenges. These results reveal that open-domain video shot retrieval is still a critical capability that multimodal large models have yet to overcome.

1 Introduction

In recent years, the rapid advances in LLMs (Yang et al., 2025; Yu et al., 2025a; Zhang et al., 2025; OpenAI, 2024; DeepMind, 2023; Anthropic, 2024) have substantially improved their capability in information retrieval. Text-based LLMs can already interact with search engines via reasoning and tool use to accomplish complex text retrieval tasks (Jin et al., 2025; Wu et al., 2025a,b; Tao et al., 2025; Yu et al., 2025b). Meanwhile, multimodal large language models (MLLMs) have begun to show the potential to perform cross-modal retrieval through reasoning and tool invocation, exemplified by methods (Geng et al., 2025). However, existing research has largely focused on text or static multimodal retrieval, while systematic investigation of video retrieval remains limited, despite its richer temporal structure and more complex semantics.

In practical video editing workflows, videos are typically divided into multiple basic units—shots (Yang et al., 2023). A shot generally refers to a continuous sequence of frames between one cut or transition and the next. As the fundamental building blocks of a video, shots not only constitute the temporal axis of the video but also serve as important carriers of content expression (Lu et al., 2025). When video editors attempt to obtain a specific shot, they often need to preview the footage in advance and remember the timestamps, which makes video editing largely dependent on personal experience and raises the entry barrier for newcomers.

Against this backdrop, a natural question arises: can current MLLM interact with search engines through reasoning and tool use to accomplish shot-level video retrieval and temporal localization? Answering this question is not only of academic importance but also directly relates to the practical potential of real-world video editing applications.

To this end, we propose ShotFinder, a shot retrieval benchmark tailored to realistic video editing workflows. ShotFinder systematically formulates the core requirements of video editors in footage retrieval as shot descriptions and five single-factor constraints in terms of temporal, color, style, audio, and resolution (Pardo et al., 2021; Bonneel et al., 2013; Tian et al., 2018; Tu et al., 2021). While keeping the task controllable, ShotFinder enables analysis of how different descriptive factors affect video retrieval and temporal localization. To mitigate biases introduced by content distribution, we adopt a constraint-aware topic assignment strategy and construct a dataset containing 1,210 high-

quality video shots through LLM-based generation with human verification. For evaluation, we employ a keyframe-based LLM-assisted assessment method, which more effectively captures performance differences among methods under complex conditions. The difference between ShotFinder and previous benchmarks is shown in Figure 1.

Building upon the benchmark, we further propose ShotFinder, a text-description-driven method for open-domain video shot retrieval and localization. Given a shot description and optional additional constraints, the method follows a staged processing pipeline. First, it performs query expansion based on video imagination, lifting shot-level descriptions to video-level search semantics. It then interacts with search engines to collect a set of candidate videos from online video platforms. Finally, it conducts description-guided temporal localization to identify the target shot that matches the input semantics. By organically integrating shot-level understanding, open-domain video retrieval, and description-driven fine-grained localization.

Based on experiments on ShotFinder benchmark, we conduct a comprehensive evaluation of human performance as well as a variety of closed-source and open-source models. The key findings are summarized as follows: (1) **Closed-source models perform better overall but still fall far short of human-level performance.** Among the closed-source models, GPT-5.2 achieves the highest average accuracy, showing clear advantages in **Temporal** and **Style**, while still being limited in **Color**. (2) **Open-source models are competitive on certain subtasks.** The Qwen3 series achieve strong results, especially in **Temporal** and **Color**, but still lags behind closed-source models in **Style**. (3) **There is a clear imbalance across task categories.** **Temporal** is the best-performing category overall, whereas **Color** and **Style** remain the major challenges for current models. (4) **Model scale does not guarantee consistent gains across all tasks.** Qwen3-Omni-30B-A3B performs close to GPT-5.2 on **Temporal**, suggesting that architectural design and multimodal alignment strategies play a critical role beyond model scale alone.

In summary, the main contributions are as follows: we introduce ShotFinder, the first benchmark for open-domain video shot retrieval, and propose an imagination-driven retrieval method. Experiments on the ShotFinder benchmark reveal the key challenges faced by current models in this field and point to directions for improvement.

2 Related Works

Web-Based Information Retrieval Web-Based Information Retrieval involves utilizing web resources such as search engines, web pages, and external tools to autonomously acquire and synthesize information. Systems like Search-R1 combine search engine queries with reasoning abilities, continuously refining answers through dynamic search interactions. WebDancer adopts a curriculum-driven strategy, guiding the agent through web-based tasks to perform multi-step reasoning and collect information from multiple sources. WebWatcher further extends this by integrating tools such as text and image search, code interpretation, and webpage browsing, enabling the agent to handle more complex multi-step queries. Despite these advancements, research in this area remains primarily focused on text or static multimodal information retrieval, with limited systematic exploration of video shot retrieval tasks that involve temporal structure and semantic complexity. To address this gap, we propose a benchmark for video shot retrieval in real-world editing scenarios, which includes 6 tasks and covers 20 topics.

Video Temporal Grounding Currently, there has been significant progress in Video Temporal Grounding (VTG) tasks for both short and long videos. For short videos, methods like Moment-DETR (Lei et al., 2021) effectively predict moments using a transformer-based architecture, enabling precise localization of key moments. Non-DETR methods, such as UMT (Liu et al., 2022) and Mr.BLIP (Rodriguez et al., 2025), further enhance the accuracy of moment prediction in short videos by leveraging multimodal cues and large language models that combine visual and textual information. For long videos, CONE (Hou et al., 2023) adopts a coarse-to-fine alignment framework with a hierarchical pipeline to handle the sparsity of key moments. However, these methods typically assume that the video already exists and only perform tasks based on known video content, which limits their application when the video has not yet been obtained. To address this limitation, we propose a new method that combines VTG with video retrieval. This approach not only resolves the issue of retrieving relevant videos but also extends temporal grounding to a new domain, allowing the model to perform temporal localization within the retrieved videos, thus advancing the integration of video retrieval and VTG tasks.

Figure 1: Comparison of ShotFinder benchmark with (a) text benchmarks and (b) other multimodal benchmarks. ShotFinder targets open-domain video retrieval. It requires the model to perform "Video Imagination" to bridge the gap between shot descriptions and full video search, followed by searching for URLs and grounding shots.

3 ShotFinder Benchmark

3.1 Task Definition

This paper proposes an open-domain shot retrieval task: given a fine-grained natural language description corresponding to the key frame of the target shot (which may include subject/action/scene information, and possibly constraints such as temporal sequence, color, style, audio, resolution, etc.), the system is required to search the internet and return video shots that are semantically consistent with the description and satisfy the given constraints. This task faces the following unique challenges:

1. **Information imbalance**. The input consists of partial information of the key frame of the shot, while the search target is the entire video. This is equivalent to using a small portion of the video's information to match the full video. As a result, it may happen that the overall theme of the video is close, but the key moment described in the query is missing.

2. **Information extraction and compression**. The descriptions are usually very detailed and long, making them unsuitable for direct retrieval queries. The model needs to extract searchable key entities/attributes/actions from the description, distinguish between strong constraints and weak descriptions, and compress them into effective retrieval representations (e.g. keywords).

3. **Video redundancy and long context pressure**. Videos contain a large number of irrelevant frames and multiple event segments. The system must both locate the matching moment amidst redundant content and model the temporal clues. This presents higher requirements for frame sampling, cross-frame aggregation, and the context length that the model can handle.

3.2 Data Collection

3.2.1 Task Design

This classification system is derived from real-world editing retrieval scenarios: when editors search for footage, the basic need is "whether the content of the shot meets the description," while higher-level needs involve "whether the shot has a specific narrative function or audiovisual presentation effect." To address this, we divide the retrieval task into shot descriptions and factor constraints, with the latter introducing additional constraints on top of the shot descriptions.

Shot Description Shot Description refers to retrieval tasks where the candidate shot is only required to match the description of the shot. The description mainly includes attributes of people/objects, actions, and scene elements, and the judgment is based on whether the visual content is consistent. For example: "An older male with silver-gray, center-parted long hair, darker skin tone, wearing a dark high-collared suit, giving a speech with full attention." In this setting, as long as the key frame of the candidate shot semantically matches the above content, it is considered a match.

Shot Description + Single Constraint In real-world editing, the selection of "usable footage" depends not only on "what's in the shot" but also on the shot's position in the narrative, overall atmosphere, or audiovisual style. To controllably analyze the impact of these factors on retrieval, we do not directly cover all the complex combinations of narrative constraints. Instead, we adopt a design where "only one additional factor is added to shot

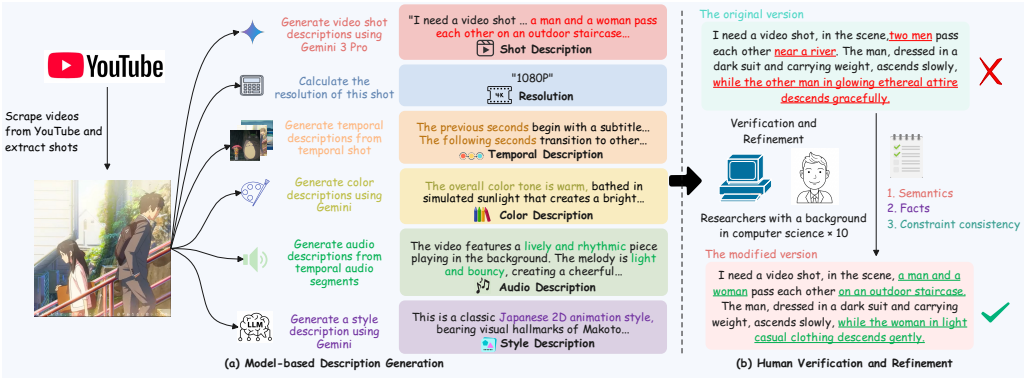


Figure 2: ShotFinder construction pipeline. (a) **Model-based Description Generation**, where Gemini generates shot descriptions with five specific constraints (Resolution, Temporal, Color, Style, Audio); and (b) **Human Verification and Refinement**, where experts refine the data to ensure semantic accuracy and constraint consistency.

description." This extends the task to the following five single-factor setups:

These constraints include **Temporal**, which require the shot to align with conditions related to preceding or following shots or action phases. We also consider **Color**, where the overall tone and warmth/coolness of the shot must match the description, capturing the atmospheric differences that color can convey. Additionally, **Style** are included, where the visual style, such as live-action footage or 3D/2D animation, must align with the description to reflect the impact of style on perception. **Audio** are also introduced, requiring the presence of specific sound elements or audiovisual relationships, such as speech, background music, or environmental sounds, to introduce cross-modal constraints. Finally, **Resolution** are applied, ensuring that technical attributes like clarity meet the requirements outlined in the description, reflecting limitations on the usability of the footage.

Therefore, our research includes 6 task settings. Shot Description, and Shot Description with the additional constraints of (**Temporal**, **Color**, **Style**, **Audio**, **Resolution**).

Why focus only on "Single Constraint" and not multi-factor combinations? The primary limitation of our study is based on methodological and experimental interpretability considerations:

Control and Attribution If multi-factor combinations were directly introduced, performance changes would be difficult to attribute to any single constraint. A single-factor design can more clearly answer "What challenges does retrieval face after introducing a specific constraint, and how much does performance change?"

Repeatability of Evaluation and Clear Boundaries The matching standards for single-factor constraints are easier to strictly define, thereby reducing subjectivity and improving reproducibility.

Layered Complexity Multi-constraint shot retrieval is a more complex constraint problem, which we regard as a future research direction. Currently, by breaking it down into single factors, we aim to establish a clear standard, providing a benchmark and methodology for future research.

3.2.2 Topic Allocation Strategy

Different constraint factors exhibit significant differences in their semantic and perceptual adaptability to video content. For example, **Color** typically relies on clear visual appearance and color distribution, while **Audio** focuses more on audio cues and their semantic information. If a unified set of topics is forced across all tasks, it could introduce many samples that are weakly related or even mismatched with specific constraints, thereby weakening the analysis of the influence of the constraints and potentially introducing additional noise.

Based on these considerations, we adopt a constraint-aware topic allocation strategy. Specifically, for each type of constraint, we select a topic set that is highly relevant to its semantic characteristics, and within that set, we aim to cover as diverse a range of video content forms as possible. At the same time, we deliberately avoid establishing a one-to-one correspondence between topics and specific constraints, achieving weak coupling between constraints and video content, thereby reducing the bias introduced by the content itself.

In selecting video topics, we refer to the content classification of the mainstream video platform Youtube. We ultimately selected 20 common video topics, covering genres such as Variety Shows, Animation, Dance, Tech, Food, Automotive, Sports, Lifestyle Vlogs, Film, TV series, Documentary, Gaming, Music, Knowledge, Fashion, Animals, Visual Arts, Tourism, Fitness, and Parenting.

In practical implementation, Shot Description does not impose additional constraint factors, so they cover all topic categories to ensure the broadness of content distribution. For tasks with added constraint factors, we select the appropriate topic set based on the type of constraint: **Temporal** mainly corresponds to topics with clear event structures or action flows, such as knowledge, fitness, food, sports, automobiles, games, movies, and TV dramas; **Color** focuses on topics with significant visual features, such as parenting, fashion, technology, art, documentaries, and travel; **Style** covers content types with prominent style attributes, such as variety shows, animation, games, and music; **Audio** primarily involves topics with rich audio information, such as TV dramas, documentaries, dance, and animals; and **Resolution** corresponds to content types that are more sensitive to visual clarity, such as fashion, movies, lifestyle vlogs, and fitness. For detailed classification rules, see Appendix A.

3.2.3 Description Generation Pipeline

Model-based Description Generation As shown in Figure 2(a), to efficiently construct a large-scale and structurally consistent description dataset, we crawl videos from YouTube and use large models for automated description generation. In this phase, we use Gemini-3-Pro to generate Shot Description, as well as various constraint descriptions (such as **Temporal**, **Color**, **Style**, **Audio**, and **Resolution**). The generated descriptions provide initial candidate texts for subsequent manual verification. The specific process details and generation templates are provided in Appendix C.

Human Verification and Refinement As shown in Figure 2(b), to ensure the accuracy and consistency of the data, we introduce a human verification and refinement phase. Specifically, we recruited 10 researchers with backgrounds in computer science to review each of the model-generated descriptions manually. During the verification process, annotators were required to assess the correctness of the descriptions in terms of semantic, factual, and constraint consistency, and to revise any errors or imprecise statements they identified. Through this manual correction and revision process, we effectively eliminated the noise introduced during the automated generation phase, ultimately constructing a dataset of 1,210 high-quality samples.

3.3 Dataset Statistics

As shown in the Figure 5, the benchmark dataset consists of 1,210 samples, evenly distributed across six tasks. These tasks cover a diverse range of 20 topics, ensuring broad content representation. The videos used for this dataset span a wide range of durations, reflecting a variety of video types. This distribution provides a comprehensive foundation for evaluating the model's performance across different domains, constraints, and content forms.

3.4 Evaluation Metrics

We first attempted to use CLIP-based (Radford et al., 2021) similarity metrics to evaluate the consistency between the model's retrieval results and the target shots. However, in preliminary experiments, we found that this type of global visual-text embedding metric struggled to effectively capture fine-grained description constraints. The evaluation results showed significant discrepancies with human judgment. Relevant analysis and comparison results are provided in Appendix B.

Based on these observations, we further adopted large-model-assisted evaluation as the primary evaluation method. Specifically, considering the cost and complexity of directly evaluating video shots, we use keyframes of the video shots as an approximate representation of the shot content. For each retrieval result, we input the corresponding keyframe, the input textual description, and the ground-truth frame into MLLM, which then determines whether the retrieved video shot meets the given description in terms of both semantics and constraints. The evaluation prompt is in Appendix D.

4 ShotFinder Method

4.1 Overview

ShotFinder is a video shot retrieval and localization process based on textual descriptions. Given a shot description (and optional additional constraints), this method performs three stages: query expansion, candidate video retrieval, and description-based temporal localization, as shown in Figure 3. First, the Generator module generates video-

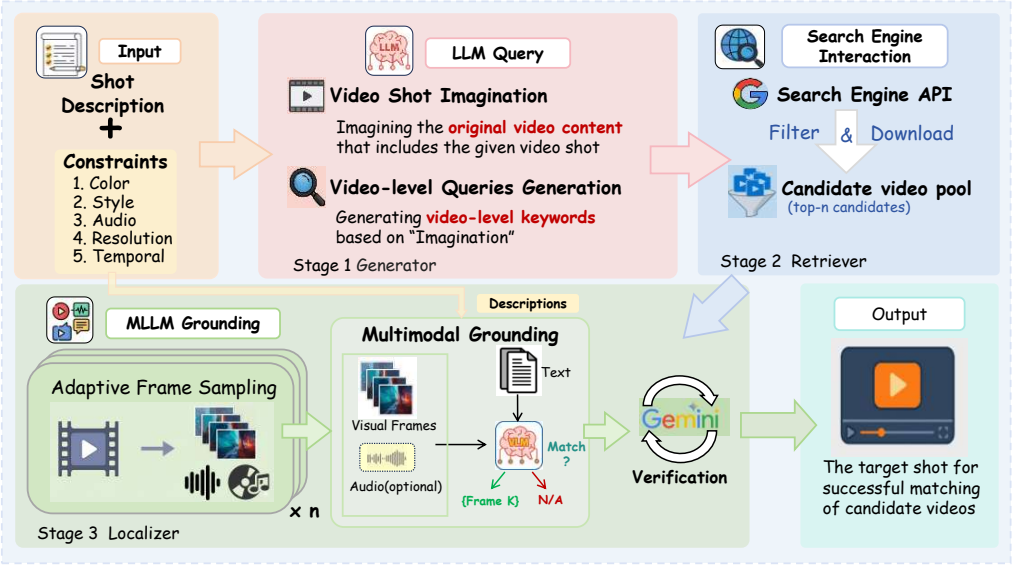


Figure 3: Illustration of the ShotFinder method pipeline. (1) **Generator**, utilizing "Video Imagination" to expand shot descriptions into effective search queries; (2) **Retriever**, fetching and filtering candidate videos from the web; and (3) **Localizer**, employing MLLMs with adaptive frame sampling to precisely locate the target shot.

level search queries through a video association mechanism. Next, the Retriever module fetches a candidate video set from search engines. Finally, the Localizer module performs temporal localization of the target shot based on the description. This staged design effectively combines shot-level descriptions with open-domain video retrieval.

4.2 Generator: Query Expansion via Video Imagination

Given a video shot description (and optional additional constraint descriptions), one straightforward approach is to ask a large model to summarize the input text or extract keywords, using them as search queries. However, in preliminary experiments, we found that this strategy tends to be limited to the local semantics of the shot, making it difficult to effectively cover high-level concepts commonly found in real video titles, tags, or descriptions, resulting in low retrieval recall.

To address this, we propose a query expansion strategy based on video imagination. Specifically, we guide the model to first infer what type of complete video the shot is most likely to appear in—i.e., we ask the model to "imagine" the original video content that includes the shot (such as the video's theme), and then generate search keywords based on this imagination. Unlike text summarization, this process incorporates video-level semantics, making the query more aligned with how videos are organized and labeled on online platforms. The imagination prompt is in Appendix E.

4.3 Retriever: Web Video Retrieval

After generating the search expansion, we perform video retrieval through a search engine to construct a candidate video set. Specifically, we use the generated keywords to interact with the search engine, retrieving publicly available video links relevant to the query, and then filter out n accessible YouTube video URLs as candidate results.

To ensure the stability of subsequent processing, we retain only the links that are successfully accessible and contain complete video content, and download the corresponding video files locally for later description-based shot localization.

4.4 Localizer: Description-guided Temporal Localization

For each candidate video, we employ an adaptive frame sampling strategy based on video duration. Specifically, videos of different lengths correspond to different frame sampling strategies to ensure temporal coverage while controlling the input size. The extracted frames are used to approximate the content of the video, providing visual input for the subsequent description-based shot localization.

During the localization phase, we input the description from which the search queries are generated and the extracted frames into MLLM, and ask the model to determine which frame matches the given description. For tasks involving **Audio**, we

	Shot	Temporal	Color	Style	Resolution	Audio	Avg.
Human	85.1	91.6	91.4	83.3	91.7	87.5	88.5
Closed-source Models							
Gemini-3-Pro	22.5	31.0	15.7	26.5	21.0	**30.0**	24.4
Gemini-2.5-Pro	23.5	29.5	12.9	17.5	18.5	26.5	21.3
GPT-5.2	**25.5**	**35.5**	15.7	**32.5**	**26.0**	–	**26.9**
GPT-5-mini	12.0	21.0	10.5	17.0	16.0	–	15.2
Claude-4.0-Sonnet	18.0	30.0	8.1	15.5	19.5	–	18.1
Open-source Models							
Qwen3-Omni-30B-A3B	19.0	33.0	11.4	18.5	17.5	21.0	20.0
Qwen3-VL-235B-A22B	18.5	28.5	**20.5**	17.0	18.5	–	20.6

Table 1: Performance comparison of closed-source and open-source models on ShotFinder benchmark.

additionally provide the video's audio as input. If the model determines that no shot in the candidate video matches the description, it outputs N/A.

Finally, we input the description from which the search queries are generated, the ground-truth frame, and the frame selected by the model into MLLM, which determines whether the model has successfully localized the shot that matches the description. This two-stage localization and verification mechanism effectively reduces the uncertainty introduced by a single model judgment.

5 Experiments

5.1 Baselines and Settings

To evaluate the performance of different models, we use both closed-source and open-source models as baselines. The closed-source models include Gemini-2.5-Pro (DeepMind, 2023), Gemini-3-Pro, Claude 4.0 Sonnet (Anthropic, 2024), GPT-5.2 (OpenAI, 2024), and GPT-5-mini (OpenAI, 2024), while the open-source models include Qwen3-VL-235B-A22B-Instruct (Yang et al., 2025), and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025).

In this experiment, we consistently use Gemini 2.5 Pro for evaluation. We refer to the video frame sampling settings used in VLMEvalKit (Duan et al., 2025) and VideoMME (Fu et al., 2025). Specifically, for closed-source models, 64 frames are uniformly sampled from videos shorter than 3 minutes, 128 frames from videos between 3 and 10 minutes, and 192 frames from videos longer than 10 minutes. For open-source models, 64 frames are uniformly sampled from all videos. Each model generates two search queries, with two URLs extracted for each query. To maintain consistency, all models are accessed via API calls with the temperature set to 0.

5.2 Main Results

Table 1 presents the performance of different models on ShotFinder, with the key conclusions:

Human vs. Model Comparison Human evaluation performs the best across all categories, and the best-performing models show significant discrepancies compared to human evaluation, highlighting the challenge of the benchmark. Human evaluation can consider the details of video shots, including style, color, and other subtle details, while existing models struggle to capture these details.

Closed-source Models Among the closed-source models, GPT-5.2 performs the best with an average accuracy of 26.9%, showing strong performance in most categories, especially in **Temporal** and **Style**. However, its performance in **Color** is relatively low. Gemini-3-Pro follows closely with an average accuracy of 24.4%, excelling in **Temporal** and **Audio**, but lagging behind in **Color**. GPT-5-mini and Claude-4.0-Sonnet show lower overall performance, indicating that smaller or less specialized models struggle more in this benchmark.

Open-source Models The Qwen3 series demonstrates competitive performance, with Qwen3-VL-235B-A22B scoring an average of 20.6. While it slightly outperforms Gemini-3-Pro in **Color**, it does not surpass closed-source models in **Temporal** and **Style**. Qwen3-Omni-30B-A3B shows similar results, with an average accuracy of 20.0%, performing strongly in **Temporal** and **Audio**.

Performance Trends **Temporal** is the best-performing category across all tasks. **Color** is generally lower, but the gap between closed-source and open-source models in these categories is not significant. **Style** shows a larger split, with models like GPT-5.2 and Gemini-3-Pro having significantly stronger capabilities.

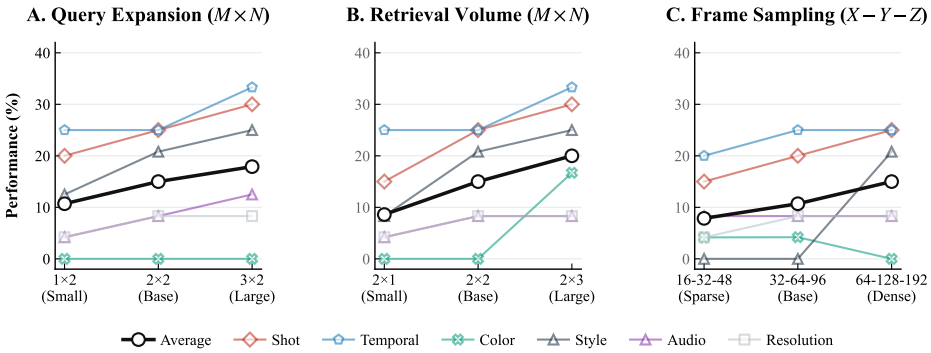


Figure 4: Results of further analysis. **Notation** ($\mathbf{M} \times \mathbf{N}$) : M = Search Queries, N = Candidate URLs per Query. **Sampling** ($\mathbf{X} - \mathbf{Y} - \mathbf{Z}$) : Frames sampled for video durations of < 3 min, $3 - 10$ min, and > 10 min respectively.

Model Scalability Larger models do not necessarily lead to better performance across all categories. Qwen3-Omni-30B-A3B performs close to GPT-5.2 on **Temporal**. Both GPT-5.2 and Gemini-3-Pro show stable performance with their larger architectures, but smaller open-source models still deliver reasonable results.

5.3 Further Analysis

In this section, we randomly selected a subset of the data for further analysis. Detailed further analysis data information can be found in Appendix F.

Effectiveness of the number of search queries. As shown in Figure 4A, increasing the number of queries (from 1×2 to 3×2) significantly improves the model's performance, especially in **Shot** and **Style**. This indicates that increasing the number of queries helps bridge the granularity gap between shot-level descriptions and video-level metadata. Although the performance improvement in **Color** is relatively small, **Temporal** and **Style** show clear benefits with an increase in the number of queries, suggesting that additional queries can better capture complex temporal and style features.

Effectiveness of the number of downloaded videos. As shown in Figure 4B, increasing the number of videos downloaded per query (from 2×1 to 2×3) improves the model's overall performance, particularly in **Shot** and **Audio**. This suggests that more candidate videos provide additional matching information, especially for features that are difficult to obtain directly through simple descriptions. However, the improvement in **Resolution** is relatively slow, indicating that its improvement is related to the inherent quality of the video, rather than the number of retrieved candidates.

Effectiveness of the number of frames extracted. As shown in Figure 4C, increasing the number of extracted frames consistently improves overall performance. The gains are more pronounced on **Shot** and **Temporal**, indicating that denser temporal sampling helps the model better capture shot boundaries and ordering. In contrast, improvements on **Color** and **Style** remain limited, suggesting that increasing frame density alone is insufficient to resolve fine-grained visual attributes.

Failure cases. In this study, we analyzed common pipeline errors. Generator errors occur when the MLLM misinterprets input, causing incorrect search expansions and irrelevant video retrieval. Partial search expansions can also omit crucial information, leading to irrelevant or partially relevant videos and inaccurate shot localization. Retriever errors happen when the MLLM generates accurate search expansions, but the retrieval stage either fails to find valid videos or retrieves irrelevant ones, leading to incorrect frame extraction. Localizer errors arise when, despite accurate search expansions and relevant video retrieval, issues with frame sampling lead to incorrect frames being extracted, or when the model mistakenly issues a judgment that contradicts human annotations. A more detailed analysis can be found in Appendix G.2.

6 Conclusion

In this paper, we propose ShotFinder, a benchmark for open-domain video shot retrieval in realistic video editing scenarios and an imagination-driven retrieval method. ShotFinder benchmark evaluates MLLMs based on shot descriptions and five practically meaningful constraint dimensions. Through MLLM evaluations, we identify significant capability gaps both among existing models and between models and humans, and reveal the challenges faced by current MLLMs in video shot retrieval tasks as well as their optimization directions.

Limitations

This study has several limitations. (1) In this study, we used key frames as a substitute for the full video shots, considering the cost and complexity of working with complete video clips. Future work will explore better methods for handling entire video shots. (2) Currently, we have focused on the impact of single-factor constraints on video editing. Multi-factor narrative editing, which is a more complex and realistic challenge, requires addressing combined constraints in a holistic manner, and we plan to explore solutions for this in future research. (3) For query generation, we employed single-turn interactions, but we aim to expand to multi-turn interactions in future iterations. After retrieving videos from search engines, we applied basic usability filtering, but we intend to implement more refined filtering based on HTML metadata from webpages for deeper content assessment. (4) In terms of frame extraction, we used an adaptive sampling strategy based on video length but did not consider the impact of video content. Future work will investigate more suitable frame extraction algorithms that better account for content-specific factors. (5) Due to cost constraints and task complexity, we only evaluate a limited set of state-of-the-art models with relatively strong capabilities. As more powerful models become available in the future, we plan to expand the range of evaluated models accordingly.

Ethical Considerations

In this study, we only crawl publicly available, non-infringing video content from the YouTube platform and adhere to the platform's terms of service and privacy policy. All data is used solely for academic research, with strict measures to ensure anonymity and privacy protection. The models used in this study are accessed via API calls, ensuring compliance with the terms of service set by the respective API providers. The models themselves were neither distributed nor modified in any way, ensuring that no intellectual property rights were violated.

References

Anthropic. 2024. Claude 3 models. https://www.anthropic.com/index/introducing-claude-3. Accessed: 2025-09-21.

Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. 2013. Example-based video color grading. *ACM Trans. Graph.*, 32(4):39–1.

Constantinos K Coursaris, Sarah J Swierenga, and Ethan Watrall. 2008. An empirical investigation of color temperature and gender effects on web aesthetics. *Journal of usability studies*, 3(3):103–117.

Google DeepMind. 2023. Gemini: Google's multimodal models. Https://deepmind.google/technologies/gemini/.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Xiaozhe Li, Chaoyou Fu, and 13 others. 2025. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 1792–179210. IEEE.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wing-Kwong Chan, Chong-Wah Ngo, Zheng Shou, and Nan Duan. 2023. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *Preprint*, arXiv:2209.10918.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei

Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *Preprint*, arXiv:2107.09609.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. *Preprint*, arXiv:2203.12745.

Chen Yi Lu, Md Mehrab Tanjim, Ishita Dasgupta, Somdeb Sarkhel, Gang Wu, Saayan Mitra, and Somali Chaterji. 2025. Skald: Learning-based shot assembly for coherent multi-shot video creation. *Preprint*, arXiv:2503.08010.

OpenAI. 2024. Gpt-4 technical report.

Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. 2021. Learning to cut by watching movies. *Preprint*, arXiv:2108.04294.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. *Preprint*, arXiv:1608.02192.

Hector Rodriguez, Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. 2025. Chrono: A simple blueprint for representing time in mllms. *Preprint*, arXiv:2406.18113.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. Webshaper: Agentically data synthesizing via information-seeking formalization. *Preprint*, arXiv:2507.15061.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. *Preprint*, arXiv:1803.08842.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. Webdancer: Towards autonomous information seeking agency. *Preprint*, arXiv:2505.22648.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025b. Webwalker: Benchmarking llms in web traversal. *Preprint*, arXiv:2501.07572.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Guoxing Yang, Haoyu Lu, Zelong Sun, and Zhiwu Lu. 2023. Shot retrieval and assembly with text script for video montage generation. ICMR '23.

Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, Yibo Yan, Tianlong Xu, Qingsong Wen, Zhang Zhang, Yan Huang, Liang Wang, and Tieniu Tan. 2025a. Aligning multimodal llm with human preference: A survey. *Preprint*, arXiv:2503.14504.

Tao Yu, Zhengbo Zhang, Zhiheng Lyu, Junhao Gong, Hongzhu Yi, Xinming Wang, Yuxuan Zhou, Jiabing Yang, Ping Nie, Yan Huang, and Wenhu Chen. 2025b. Browseragent: Building web agents with human-inspired web browsing actions. *Preprint*, arXiv:2510.10666.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025. Mm-rlhf: The next step forward in multimodal llm alignment. *Preprint*, arXiv:2502.10391.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. Towards automatic learning of procedures from web instructional videos. *Preprint*, arXiv:1703.09788.

A Theme Classification

Temporal To strictly define the ratio of Pre-context, we selected categories such as Knowledge, Fitness, and Food, as these topics inherently have a strong process-oriented nature. As described by (Zhou et al., 2017) in their research on learning steps through instructional videos, such content relies on initial "ingredients" or "instructions" to build subsequent temporal logic; viewers must understand the prerequisites to grasp the core content. For Post-context categories, we chose Sports,

Automotive, and Gaming. Based on (Giancola et al., 2018) research on action recognition in sports videos, the semantic focus of such content often shifts to "event outcomes" (e.g., goals, win/loss status, race rankings), making their temporal grounding dependent on final discrete results rather than initial states. Movies and TV Series are isolated in the Pre-context and Post-context category because they represent complex narrative structures, requiring both setup and resolution, and this bidirectional temporal dependency aligns with the narrative coherence required in narratology. For data distribution, the ratio of Pre-context, Post-context, and both is 75:75:50, with topics in each category distributed equally.

Color To create clear color boundaries, we applied strategic sampling: we mapped Parenting and Fashion to warm tones, utilizing the family warmth in Parenting content and the warm lighting in Fashion videos, which statistically form a significant contrast with cool-toned backgrounds; Tech and Art were assigned to cool tones because technology and digital art images frequently use cool hues (blue/cyan), high-saturation composite colors, and metallic finishes, enhancing the semantics of "precision," "technological maturity," and "futurism"—cool tones create a perception of technology and objectivity visually (Coursaris et al., 2008), making them widely used in product illustrations and digital art; Finally, we used Documentary and Travel as neutral categories, retaining natural light distribution from the real world and avoiding overly stylized color grading, serving as the true visual baseline for the dataset. For data distribution, the ratio of warm tones, cool tones, and neutral is 70:70:70, with topics in each category distributed equally.

Style This choice achieves strict balanced distribution by assigning a single representative category to each quadrant. Variety Shows are the primary Live-action component, characterized by standard studio shooting techniques; Animation is explicitly defined as 2D Animation; Gaming represents 3D Animation/Rendering, as discussed by (Richter et al., 2016), where modern game engines generate realistic images but differ fundamentally in texture mapping and lighting calculations from real cameras, helping the model distinguish between passive CGI and real-time rendering. Lastly, Music was chosen as the representative of Pure Graphics, particularly for lyric videos and dynamic typography, where high-density overlay text and dynamic graphics create unique visual features, ensuring the model can effectively recognize and process non-natural scene content dominated by text elements and user interfaces. For data distribution, the ratio of live-action, 2D animation, 3D animation, and pure graphics is 50:50:50:50, with topics in each category distributed equally.

Audio To satisfy the 100:50:50 ratio, we prioritized Human Voice, selecting TV Series and Documentaries because these dialogue-driven genres provide the most continuous speech signals. This aligns with (Gemmeke et al., 2017), whose AudioSet ontology identifies "Human Voice" as the core, essential explicit signal for video semantic understanding. Dance was selected to represent Music because its visual choreography is strictly synchronized with the audio rhythm, offering superior multimodal alignment compared to mixed audio types. Finally, Animals was chosen to represent Ambient Sound, as this content is mainly filmed in natural, outdoor environments, ensuring that the audio track is naturally dominated by biological calls and environmental textures (e.g., wind, water sounds), rather than human noise or dialogue. For data distribution, the ratio of human voice, music sound effects, and ambient sound is 100:50:50, with topics in each category distributed equally.

Resolution We built a 100:100 (1:1) split between professionally produced content and user-generated content (UGC). Fashion and Movies serve as anchors at 1080p (High-Res), as these genres inherently prioritize extreme visual fidelity and complex color grading to provide an immersive viewing experience, providing the model with clean, artifact-free reference samples. In contrast, Life Vlogs and Fitness are mapped to 720p (Low/-Variable Res) to capture "outdoor" authenticity; this aligns with (Hosu et al., 2020), which emphasizes the necessity of learning from real distortion. These categories introduce typical mobile filming artifacts such as focus errors and compression artifacts, training the model to robustly handle the diverse quality spectrum found in real-world social media streams. For data distribution, the ratio of 1080p and 720p is 100:100, with topics in each category distributed equally.

B Model evaluation vs. human evaluation

As shown in Figure 6, we present the confusion matrices between CLIP evaluation and human evaluation, as well as between MLLM evaluation and human evaluation. From the figure, it is evident that there is a significant discrepancy between CLIP evaluation and human evaluation, while MLLM evaluation aligns closely with human evaluation in most cases. This indicates that, in this experiment, the CLIP evaluation method is not suitable, whereas MLLM's judgments demonstrate strong reliability and can effectively replace human evaluation. We provide a case study in Appendix G.1. In our experiments, we set the CLIP similarity score threshold to 0.7.

C Description Generation

C.1 Data Acquisition

The benchmark dataset is collected from YouTube, covering six task categories designed to evaluate multimodal understanding capabilities in video editing scenarios. The data acquisition process varies based on task requirements, with different preprocessing and filtering strategies applied to different task types.

Task Categories The dataset includes six task types: shot detection, temporal grounding, resolution classification, color classification, style classification, and audio description. Each task targets a specific aspect of video editing tasks.

Standard Tasks (Direct Collection) For shot detection, temporal grounding, and resolution classification tasks, video segments are directly collected through keyword-based YouTube search and undergo standard preprocessing without additional filtering requirements. These tasks focus on structural and technical attributes that can be reliably extracted through automated processing.

Filtered Tasks (Dual-Model Strategy) For color classification, style classification, and audio description tasks, video segments require specific characteristics that cannot be directly obtained through metadata or keyword search. Since these attributes are intrinsically linked to video content rather than searchable metadata, we employ automated filtering strategies to ensure data quality.

Dual-Model Filtering for Color and Style For color and style classification tasks, we implement a parallel dual-model filtering mechanism using both Qwen-2.5-VL-72B and Gemini-3-Pro to ensure annotation consistency and reduce model bias. Both models independently analyze the same video frames, and they must reach agreement on the classification results meeting the requirements for them to be considered valid. The system automatically retries up to 3 times if either model returns unknown or ambiguous results. Only frames where both models agree on the same category are retained in the dataset. This consensus-based approach significantly improves filtering accuracy and ensures consistent annotation quality.

Automated Audio Filtering For the audio description task, we use Gemini-3-Pro with an automated audio quality assessment system to filter video segments. The system analyzes whether the audio meets the requirements of our task, and only segments that meet the criteria are retained for annotation.

Preprocessing Pipeline After passing the filtering stage (for color, style, and audio tasks) or direct collection (for shot, temporal, and resolution tasks), all video segments undergo unified preprocessing. For each video, we randomly select a target timestamp and extract a 10-second window centered on that moment. The preprocessing extracts the following multimodal assets:

Multimodal Asset Structure The extracted assets for each video segment include:

- **Target Frame:** A single keyframe (JPG format) representing the moment to be described.
- **Preceding Frames:** Three frames sampled at 1-second intervals before the target frame, providing temporal context of what happened before.
- **Following Frames:** Three frames sampled at 1-second intervals after the target frame, capturing the subsequent narrative development.

- **Audio Segment:** An MP3 file extracted from the 10-second window centered on the target timestamp.

- **Metadata:** A JSON file recording the target timestamp, video duration, resolution, and file paths of all extracted assets.

Frame Similarity Filtering To ensure temporal diversity and avoid redundant frames, we employ a CLIP-based similarity filtering mechanism. For each 7-frame sequence (target frame + 3 preceding + 3 following), we compute cosine similarity between all 9 pairs of adjacent frames using CLIP embeddings. If all pairwise similarities exceed 0.8, indicating insufficient temporal variation, the entire sequence is discarded and a new target timestamp is randomly selected. This process repeats until a sequence with adequate temporal diversity is obtained. This strategy ensures that the extracted frames capture meaningful temporal changes rather than static scenes.

C.2 Description Generation Prompts

C.2.1 Filtering Prompts

Color Filtering Prompt

Input: Target keyframe (JPG image)

System Prompt: You are a visual analysis expert specializing in color tone classification.

Task: Analyze the provided image and classify its overall color tone into one of three categories.

Categories:
Warm Tone: The image is dominated by warm colors such as orange, yellow, or red.
Cold Tone: The image is dominated by cold colors such as blue, cyan, or purple.
Neutral Tone: The image has balanced white balance with no obvious color bias.

Output: Return only one word indicating the classification result.

Style Filtering Prompt

Input: Target keyframe (JPG image)

System Prompt: You are a visual analysis expert specializing in visual style classification.

Task: Analyze the provided image and classify its visual style into one of four categories.

Categories:
Real: Live-action or cinematic footage, content based on reality.
2D Animation: Flat animation including anime, cartoons, or hand-drawn content.
3D Animation: 3D animation, game graphics, or rendered scenes.
Graphic: Pure graphics such as charts, text, or software interfaces.

Output: Return only one option indicating the classification result.

Audio Filtering Prompt

Input: Audio segment (MP3 file)

System Prompt: You are an expert in audio description classification.

Task: Analyze the provided audio clip and classify its overall audio description into one of the three categories.

Categories:
Human Voice: The audio clip primarily contains human speech or dialogue.
Background Music: The audio clip primarily contains music or background tracks, with no significant human voice.
Ambient Sound: The audio clip primarily contains natural or environmental sounds, such as wind, water, animal sounds, etc.

Output: Return only one word indicating the classification result.

C.2.2 Frame Description Prompt

Multimodal Description Generation Prompt

Input: Target keyframe (JPG image), preceding frames (3 images), following frames (3 images), and audio segment (MP3)

System Prompt: You are a professional video content analyst. Generate comprehensive multimodal descriptions for video segments.

Task Requirements:

Main Frame Description (240-260 characters): Provide a detailed description of the key frame, including subjects, actions, props, text overlays, lighting characteristics, and camera language. The description should be rich in visual details and maintain the specified character count.

Temporal Context:
Preceding Seconds: Describe what happens in the few seconds before the key frame, focusing on actions and scene transitions.
Following Seconds: Describe what happens in the few seconds after the key frame, emphasizing narrative continuity.

Color Analysis: Analyze the overall color tendency, color intensity, and lighting characteristics of the frame. Describe the dominant color palette and its emotional impact.

Style Analysis: Classify the visual style (real, 2D animation, 3D animation, or graphic) and describe its characteristic features, such as rendering techniques, artistic style, or production quality.

Audio Description: Analyze the audio track, including speech content, sound events, background music, and ambient audio. Describe the audio atmosphere and its relationship to the visual content.

Output Format: Generate descriptions in natural language for each component listed above.

C.2.3 JSON Validation Prompt

Data Validation and Formatting Prompt

Input: Raw multimodal descriptions (text format)

System Prompt: You are a professional data annotator. Organize the input raw video analysis data into strict JSON format.

Task Requirements:

Deduplication and Cleaning: Remove descriptions of non-content elements such as watermarks, progress bars, and corner icons.

Privacy Compliance: Remove specific real-world names and locations (unless they are well-known landmarks), replacing them with generic descriptions.

Bilingual Output: All description fields must generate both Chinese (_ch) and English (_en) versions.

Character Count Requirement (Critical): The segment_description_ch must be maintained at approximately 240-260 characters. The detailed content from raw_main must be fully preserved, with only privacy cleaning applied, without reducing word count.

Context Description Format (Important): The context_description_ch and context_description_en must be arrays containing two elements in the format: [description of preceding seconds, description of following seconds].

Format Requirements (Critical): Output only pure JSON objects without any Markdown markers. Do not add any explanatory text. Ensure all strings are

properly escaped.

Output JSON Schema:
{
id: youtube_<video_id>,
video_link: https://...,
video_source: YouTube,
category: ...,
timestamp: YYYY-MM-DD HH:MM:SS,
resolution: 1080P/720P,
segment_description_ch: ...,
segment_description_en: ...,
context_description_ch: [before..., after...],
context_description_en: [before..., after...],
color_description_ch: ...,
color_description_en: ...,
style_description_ch: ...,
style_description_en: ...,
audio_description_ch: ...,
audio_description_en: ...
}

D Evaluation

Evaluation Prompt

Role: You are a Visual Frame Grounding Specialist. Your task is to identify the **single most relevant frame** that best matches the user's description.

Context:
The input consists of **multiple images extracted from a video**.
The images are provided in **strict chronological order**, from the beginning to the end of the video.
The **first image has frame_id = 0**, the second image has frame_id = 1, and so on.
In addition to images, you are also given the **full audio track of the video**.
The audio is **not segmented per frame** and should be used as **global temporal and semantic context**, not as a direct frame-to-audio alignment.

Temporal Information:
The video duration is **$VIDEO_DURATION$ seconds**.
A total of **NUM_FRAMES frames** are sampled uniformly from the video.
Frame i corresponds approximately to time $\mathbf{t_i = (i + 0.5) / \$NUM_FRAMES\$ \times \$VIDEO_DURATION\$}$ **seconds**.

Target Description (The event to find):
"en_memory_data"

Instructions:

1. Joint Reasoning with Vision and Audio:
Use the **text description**, **visual content of all frames**, and the **audio track** together.
Use audio cues (e.g., speech, sound events, changes in intensity) to infer **when** the described event occurs.
Use visual evidence from frames to determine **which frame best represents that moment**.

2. Scan All Frames: Carefully analyze **all** provided images in chronological order.

3. Select the Best Match: Identify the **single frame** whose visual content is **most relevant** to the target description.
If multiple frames partially match, choose the **closest in time to the inferred audio event**.
If the match is weak or ambiguous, select the **most semantically related frame**.

4. Mandatory Selection Rule:
You **must return exactly one valid frame_id**.
The frame_id **must be an integer between 0 and NUM_FRAMES - 1**.
Do **not** return "N/A", null, or multiple values.

Output Format (JSON Only):
<tool_call>{"frame_id": <integer>}</tool_call>

E Imagination

Imagination Prompt

You are an autonomous "Video Hunter Agent". Your ultimate goal is to precisely locate the original video link based on the user's vague memory (including visual, auditory, and timeline clues). You must provide **2** keywords most likely to hit the target container.

Tool Available:
`search_videos`: Simultaneously performs search on YouTube.
Parameters: `{"query": "search keywords"}`
Note: This tool automatically handles the "youtube" suffix, so you only need to provide the core keywords.

Output Format:
<think>...your deep reasoning process...</think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["your search term1", "your search term2"]}}</tool_call>

Core Constraints & Strategy:

1. Reasoning First: Before taking any action, you must think inside <think>...</think>. You must not directly call tools or output answers without prior reasoning.

2. Keyword Generation Strategy (Abductive Reasoning): You must use "Abductive Reasoning" to bridge the gap between the user's "clip description" and the likely "full video title."

Deconstruction: Extract visual anchors (OCR, specific objects), actions, and atmosphere.

Divergent Hypotheses ("Container" Logic): You must propose different paths for what the full video actually is:

Hypothesis A (Literal): The video is explicitly about the action described (e.g., a "How-to" or generic vlog).

Hypothesis B (Metaphorical/B-roll): The clip is a visual metaphor for an abstract concept (e.g., "cutting bread with a spoon" → Search keywords: "inefficiency", "wrong tools", "productivity").

Hypothesis C (Specific Source): The clip is from a specific movie, meme, game glitch, or news event (e.g., "bucket on shopkeeper head" → Search keywords: "Skyrim bucket thief").

Selection: Select the most probable hypothesis to construct a **single best English search phrase**.

Information Source Isolation: Reference ONLY English corpus.

3. Search Strategy: You MUST perform search rounds using the best English keywords derived from your strongest hypothesis.

4. Execution Workflow: Analyze the description using Abductive Reasoning & construct the best English keyword → call `search_videos`

Example:
User: "Split-screen dance video, Y2K fashion, leg warmers, electronic music"
Agent: <think>Y2K fashion + split-screen editing + dance = likely TikTok trend or K-pop cover. Keywords: style + format.</think><tool_call>{"name": "search_videos", "arguments": {"query": ["Y2K dance split screen leg warmers", "split screen clone dance phonk"]}}</tool_call>

F Further Analysis

The data distribution in the ablation experiments is the same as in the main experiment, and the data distribution is shown in Figure 7. The model used

for generating results is Gemini 3 Pro.

For **Shot**, one data sample is randomly selected from each of the 20 topics; for Temporal, 24 data samples are randomly selected with the following distribution: Knowledge, Fitness, Food, Sports, Cars, Games, Movies, and TV Shows, with 3 samples each. Specifically, Knowledge, Fitness, and Food use Pre-context; Sports, Cars, and Games use Post-context; Movies and TV Shows use both Pre-context and Post-context. For **Color**, 24 data samples are randomly selected, with the following distribution: Parenting, Fashion, Technology, Art, Documentaries, and Travel, with 4 samples each. Specifically, Parenting and Fashion use warm tones; Technology and Art use cool tones; Documentaries and Travel use neutral tones. For **Style**, 24 data samples are randomly selected, with the following distribution: Variety Shows, Animation, Games, and Music, with 6 samples each. Specifically, Variety Shows use live-action; Animation uses 2D animation; Games use 3D animation; Music uses pure graphics. For **Audio**, 24 data samples are randomly selected, with the following distribution: TV Shows, Documentaries, Dance, and Animals, with 6 samples each. Specifically, TV Shows and Documentaries use human voices; Dance uses music; Animals use ambient sounds. For **Resolution**, 24 data samples are randomly selected, with the following distribution: Fashion, Movies, Lifestyle Vlogs, and Fitness, with 6 samples each. Specifically, Fashion and Movies use 1080P; Lifestyle Vlogs and Fitness use 720P.

G Case study

G.1 Case study for verification.

Example of case A (MLLM True & Human True)

Selected frame:

Ground truth text:

I need a video clip, two male musicians stand side by side on an open expanse of sand: one plays an acoustic guitar while the other cradles his instrument, eyes closed in impassioned vocal performance. Their postures are natural and fully immersed, with microphone stands positioned before them to form a performance array. The foreground features the subjects and their instruments; the midground reveals a flat desert landscape; and the background presents softly blurred mountain silhouettes. A horizontal composition enhances spatial depth. The overall tone is natural, neutral-leaning warm, with a transparent and clean atmosphere. Colors are soft with low saturation and smooth transitions. The lighting is bright, characterized by soft natural diffused light without strong contrast or harsh shadows, resulting in a moderately low overall contrast.

Ground truth frame:

Model's prediction: True

Human's prediction: True

Example of case B (MLLM True & Human False)

Selected frame:

Ground truth text:

I am looking for a clip, this is a keyframe from a nighttime highway racing perspective, captured with an immersive first-person wide-angle lens. The foreground is dominated by the hood of a teal-modified sports car, with sharp black air vents and the right-side mirror clearly visible, enhancing the sense of driver immersion. In the midground, a white Lamborghini convertible supercar races ahead on the left, its low-slung body and prominent rear wing standing out dramatically against the night sky. Floating text at the top displays the white English title 'Smoked a Lambo going home' and the central phrase 'wait for it...' builds suspense and foreshadows an impending overtaking

climax. Environmental lighting comes primarily from distant streetlights and digital billboards, creating high-contrast highlights against the dark sky.

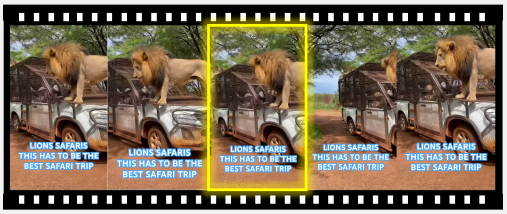
Ground truth frame:

Model's prediction: True

Human's prediction: False

Example of case C (MLLM False & Human True)

Selected frame:



Ground truth text:

I want to find a video segment, the main subject of this frame is a large, majestic adult male lion standing inside a metal transport cage mounted on the back of a truck. It has a thick, flowing mane, transitioning from golden yellow on the head to deep black around the neck, exuding strength and dignity. The lion stands tall, alert and solemn, gazing intently toward the upper right of the frame. The cage, featuring a black metal frame and dense silver wire mesh, is positioned above a lowered metal ramp and a heavy-duty white bumper with circular holes, creating a strong industrial and oppressive feel. Bright natural sunlight streams from the upper right, clearly highlighting the texture of the lion's facial muscles and mane.

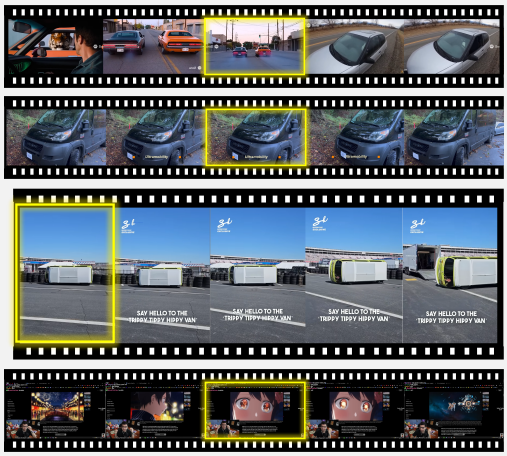
Ground truth frame:

Model's prediction: False

Human's prediction: True

Example of case D (MLLM False & Human False)

Selected frames:



Ground truth text:

I am searching for a clip, a black van drives centrally toward the camera, its headlights illuminated. The vehicle hugs the road with a low-slung profile, while the street recedes into a backdrop of urban buildings. The foreground features asphalt pavement with clear lane markings; the midground is dominated by the vehicle itself; and the background comprises multi-story buildings and street trees, resulting in a balanced, symmetrical composition.The overall tone is neutral to warm and natural, dominated by the inherent colors of the realistic scene. Color saturation is moderate to soft, providing a plain and natural visual experience. The image is bright with moderate contrast between light and shadow, featuring both highlights from direct sunlight and clear shadows cast by objects, indicating distinct light direction and dimensionality.

Ground truth frame:

Model's prediction: False

Human's prediction: False

G.2 Case study for errors occured in pipeline

G.2.1 Generator errors

In this error category, the MLLM misinterprets the input fragment, which results in the generation of incorrect search expansions. This initial error leads the web search interface to retrieve an irrelevant candidate video set, further propagating the inaccuracy to the shot localization stage. Consequently, the entire pipeline operates on irrelevant videos, ultimately yielding an erroneous final output.

Generate wrong queries

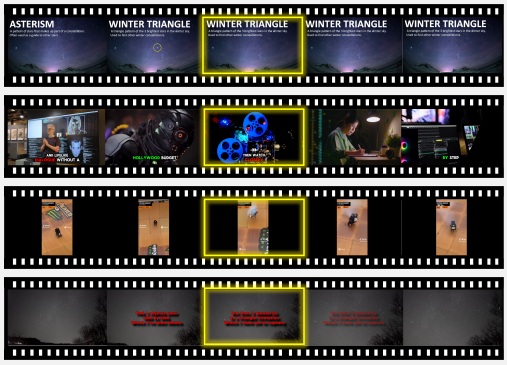
Input fragment:

Please find the link to this video: I need a video clip, three figures clad in dark gear move swiftly across the snowfield in a tight triangular formation, maintaining low postures, synchronized movements, and consistent spacing between them, casting sharp, elongated diagonal shadows. In the foreground, textured undulations of ice create intricate surface detail; the midground features the trio as the primary subjects; and the background reveals towering snow-capped peaks and glacial escarpments, establishing a composition that is expansive with pronounced depth.

MLLM output context:

<think>The description is highly detailed and specific ("three figures clad in dark gear," "triangular formation," "textured undulations of ice"), which strongly suggests it is a prompt or description of an AI-generated video, likely from OpenAI's Sora showcase. The phrasing matches the style of prompts used to demonstrate Sora's capabilities.</think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["Sora AI three figures dark gear snowfield", "three figures snow triangular formation cinematic"]}}</tool_call>

Selected frames:



Ground truth text:

I need a video clip, three figures clad in dark gear move swiftly across the snowfield in a tight triangular formation, maintaining low postures, synchronized movements, and consistent spacing between them, casting sharp, elongated diagonal shadows. In the foreground, textured undulations of ice create intricate surface detail; the midground features the trio as the primary subjects; and the background reveals towering snow-capped peaks and glacial escarpments, establishing a composition that is expansive with pronounced depth. (1080P)

Ground truth frame:

Model's prediction: False

In this type of error, the MLLM generates only partially precise search expansions, which cause the omission of crucial information from the input. This information loss leads the retrieval interface to return a candidate video set that is either partially relevant or entirely unrelated. Consequently, the system fails to perform accurate shot localization, making it impossible to extract the correct frames

for final judgment.

Generate partially correct queries

Input fragment:

Please find the link to this video: I need a video clip, a blonde woman stands centered on stage, wearing a tailored grayish-brown dress, her hands gently clasped at her chest in a poised and dignified posture, as if delivering a speech or expressing gratitude. The background features a deep blue gradient backdrop with naturally cascading folds that create soft vertical depth; the foreground subject is sharply rendered and prominently highlighted, while the midground remains free of distracting elements, resulting in a balanced and symmetrical composition.

MLLM output context:

<think>The detailed, descriptive language ("sharply rendered", "midground remains free", "deep blue gradient backdrop with naturally cascading folds") strongly suggests this is a prompt or description of a specific AI-generated video demo, likely from OpenAI's Sora or similar high-end video generation models. The visual specifics match a known Sora example.</think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["Sora AI blonde woman speech blue background ", "blonde woman grayish brown dress deep blue backdrop *(missing:handclaspe)*"]}}</tool_call>

Selected frames:

Ground truth text:

I need a video clip, a blonde woman stands centered on stage, wearing a tailored grayish-brown dress, her hands gently clasped at her chest in a poised and dignified posture, as if delivering a speech or expressing gratitude. The background features a deep blue gradient backdrop with naturally cascading folds that create soft vertical depth; the foreground subject is sharply rendered and prominently highlighted, while the midground remains free of distracting elements, resulting in a balanced and symmetrical composition.(1080P)

Ground truth frame:

Model's prediction: False

G.2.2 Retriever errors

In this type of error, the MLLM generates precise search expansions, the video retrieval stage fails to construct a valid candidate video set from the web. Due to the lack of accessible video links that match the search expansions on the search engine, the system ultimately returns a null output.

No searched videos

Input fragment:

Please find the link to this video: I need a piece of stock footage, two drummers stand side by side at the center of the stage, clad in black sleeveless tops, long trousers, and brown boots, simultaneously striking wooden war drums suspended at their waists with drumsticks. Their movements are perfectly synchronized and charged with tension. The foreground features a dark, reflective stage surface; the midground frames the performers; and the background consists of a metal mesh panel overlaid with a dot-matrix LED wall, creating a strong sense of depth. The shot is a static, eye-level medium frame that fully captures the performers' upper bodies

and their drums, emphasizing rhythmic dynamism. Overhead top lighting interplays with side-mounted green lights, sharply delineating the performers' silhouettes, while deep shadows in the background enhance the subjects' presence and evoke an intense yet enigmatic atmosphere. Real-shot. The visual style is typical of stage performance photography, emphasizing the capture of dynamic moments and realistic states. The composition presents obvious symmetry, with a realistic visual language that focuses on showcasing the on-site atmosphere and tension of movement.

MLLM output context:

<think>The user describes a specific stock footage clip featuring two synchronized drummers with war drums on a stage with green lighting and a metal mesh background. Key visual elements include the symmetry, the specific outfit (black sleeveless), and the lighting conditions. I will construct queries focusing on the drummers, the type of drum, and the lighting/stage atmosphere. </think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["two drummers synchronized war drums green lights stage ", "stock footage two drummers black sleeveless wooden drums"]}}</tool_call>

Selected frames:

No selected videos.

Ground truth text:

I need a piece of stock footage, two drummers stand side by side at the center of the stage, clad in black sleeveless tops, long trousers, and brown boots, simultaneously striking wooden war drums suspended at their waists with drumsticks. Their movements are perfectly synchronized and charged with tension. The foreground features a dark, reflective stage surface; the midground frames the performers; and the background consists of a metal mesh panel overlaid with a dot-matrix LED wall, creating a strong sense of depth. The shot is a static, eye-level medium frame that fully captures the performers' upper bodies

and their drums, emphasizing rhythmic dynamism. Overhead top lighting interplays with side-mounted green lights, sharply delineating the performers' silhouettes, while deep shadows in the background enhance the subjects' presence and evoke an intense yet enigmatic atmosphere.Real-shot. The visual style is typical of stage performance photography, emphasizing the capture of dynamic moments and realistic states. The composition presents obvious symmetry, with a realistic visual language that focuses on showcasing the on-site atmosphere and tension of movement.

Ground truth frame:

Model's prediction: False

In this type of error, thought the MLLM generates precise search expansions and the retrieval stage returns some results. However, the returned video results are all irrelevant. Consequently, the shot localization stage returns irrelevant frames.

Searched videos are all wrong

Input fragment:

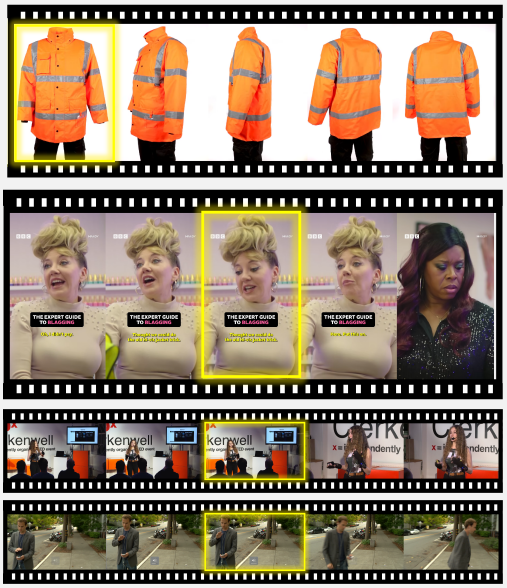
Please find the link to this video: I need a piece of stock footage, the primary subject is a standing male figure wearing an olive-green jacket, holding a luminous device in his right hand and raising his left hand in a natural, focused explanatory gesture. In the foreground lies an orange tabletop with black equipment; the midground features the main subject; and the background displays large electronic instruments alongside a blurred topographic map, establishing a clear sense of depth and layered composition. The shot is a static eye-level medium frame, tightly framing the upper body of the subject with no discernible camera movement, ensuring a stable presentation of the demonstration scene.

MLLM output context:

<think>The user describes a specific stock footage clip featuring a man in an olive-green jacket holding a luminous device with an orange table and electronic equipment. The setting suggests a tech, scientific, or military context. Keywords should focus on the visual elements and the specific action.</think>

<tool_call>{"name": "search_videos", "arguments": {"query": ["stock footage man olive green jacket luminous device orange table", "man holding luminous device explanatory gesture electronic instruments stock video"]}}</tool_call>

Selected frames from the filtered videos:



Selected frame from the dropped video (we choose the most similar frame):

Ground truth text:

I need a piece of stock footage, the primary subject is a standing male figure wearing an olive-green jacket, holding a luminous device in his right hand and raising his left hand in a natural, focused explanatory gesture. In the foreground lies an orange tabletop with black equipment; the midground features the main subject; and the background displays large electronic instruments alongside a blurred topographic map, establishing a clear sense of depth and layered composition. The shot is a static eye-level medium frame, tightly framing the upper body of the subject with no discernible camera movement, ensuring a stable presentation of the demonstration scene. (1080P)

Ground truth frame:

Model's prediction: False

G.2.3 Localizer errors

In this type of error, the MLLM successfully generates precise search expansions and retrieves relevant videos. However, the video frame sampling process is flawed, causing irrelevant frames to be extracted, which leads to a 'False' judgement.

Sample wrong frames

Input fragment:

Please find the link to this video: I need to get a clip, this keyframe captures a breathtaking moment of a massive orca leaping vertically out of an artificial pool. The orca's distinct black-and-white streamlined body is almost fully airborne, with its dorsal fin upright, showcasing immense power. As it surges upward, it generates a burst of churning white water, with droplets glistening in the sunlight. To the right, a trainer in a blue-and-black wetsuit stands at the pool's edge, facing the orca with hands raised as if applauding or giving commands, creating clear interaction. The background features a painted ocean mural, metal railings, and distant greenery, with a prominent orange traffic cone by the poolside. I vaguely recall the audio being lively and chaotic, as if I were in some sort of obstacle course game. From what I remember, there was a rapid and cheerful soundtrack, interspersed with exaggerated sound effects of objects crashing violently. I think there were also

some indistinct shouts, creating a tense yet comical and joyful atmosphere that left a lasting impression on me.

MLLM output context:

<think>The user describes a specific visual of an orca jumping near a trainer with an orange cone, combined with audio resembling a chaotic obstacle course game (likely *Wipeout* or *Fall Guys*). This suggests the user saw a meme edit. I will search for the visual elements to find the source clip and the meme context.</think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["orca jump trainer orange cone", "orca jump wipeout meme"]}}</tool_call>

Selected frame:

True frame in the same video:

Ground truth text:

I need to get a clip, this keyframe captures a breathtaking moment of a massive orca leaping vertically out of an artificial pool. The orca's distinct black-and-white streamlined body is almost fully airborne, with its dorsal fin upright, showcasing immense power. As it surges upward, it generates a burst of churning white water, with droplets glistening in the sunlight. To the right, a trainer in a blue-and-black wetsuit stands at the pool's edge, facing the orca with hands raised as if applauding or giving commands, creating clear interaction. The background features a painted ocean mural, metal railings, and distant greenery, with a prominent orange traffic cone by the poolside. (1080P)

Ground truth frame:

Model's prediction: False

In this error category, although the MLLM generates precise search expansions, a discrepancy occurs between the retrieved content and the model's judgment. Specifically, when the video retrieval stage returns an irrelevant candidate video set, the subsequent shot localization captures incorrect frames. However, the MLLM erroneously issues a 'True' judgment based on these irrelevant frames, contradicting human ground truth. Conversely, the error also manifests when the model produces a 'False' judgment despite the retrieval of correct videos and the extraction of accurate frames.

Wrong verification (I)

Input fragment:

Please find the link to this video: I am looking for a clip, this is a keyframe from a nighttime highway racing perspective, captured with an immersive first-person wide-angle lens. The foreground is dominated by the hood of a teal-modified sports car, with sharp black air vents and the right-side mirror clearly visible, enhancing the sense of driver immersion. In the midground, a white Lamborghini convertible supercar races ahead on the left, its low-slung body and prominent rear wing standing out dramatically against the night sky. Floating text at the top displays the white English title 'Smoked a Lambo going home' and the central phrase 'wait for it...' builds suspense and foreshadows an impending overtaking climax. Environmental lighting comes primarily from distant streetlights and digital billboards, creating high-contrast highlights against the dark sky.

MLLM output context:

<think>The user describes a specific POV street racing video. Key identifiers are the text overlays "Smoked a Lambo going home" and "wait for it...", along with a teal sports car chasing a white Lamborghini convertible at night. The text is the strongest search signal.</think>
<tool_call>{"name": "search_videos",

"arguments": {"query": ["Smoked a Lambo going home wait for it", "teal car racing white lamborghini night pov"]}}</tool_call>

Selected frame:

Ground truth text:

I am looking for a clip, this is a keyframe from a nighttime highway racing perspective, captured with an immersive first-person wide-angle lens. The foreground is dominated by the hood of a teal-modified sports car, with sharp black air vents and the right-side mirror clearly visible, enhancing the sense of driver immersion. In the midground, a white Lamborghini convertible supercar races ahead on the left, its low-slung body and prominent rear wing standing out dramatically against the night sky. Floating text at the top displays the white English title 'Smoked a Lambo going home' and the central phrase 'wait for it...' builds suspense and foreshadows an impending overtaking climax. Environmental lighting comes primarily from distant streetlights and digital billboards, creating high-contrast highlights against the dark sky.

Ground truth frame:

Model's prediction: True

Human's prediction: False

Wrong verification (II)

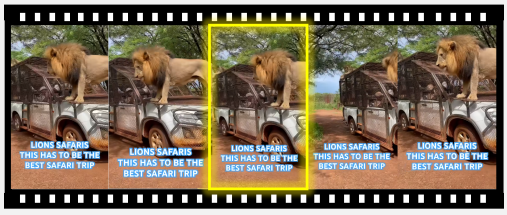
Input fragment:

Please find the link to this video: I want to find a video segment, the main subject of this frame is a large, majestic adult male lion standing inside a metal transport cage mounted on the back of a truck. It has a thick, flowing mane, transitioning from golden yellow on the head to deep black around the neck, exuding strength and dignity. The lion stands tall, alert and solemn, gazing intently toward the upper right of the frame. The cage, featuring a black metal frame and dense silver wire mesh, is positioned above a lowered metal ramp and a heavy-duty white bumper with circular holes, creating a strong industrial and oppressive feel. Bright natural sunlight streams from the upper right, clearly highlighting the texture of the lion's facial muscles and mane. I vaguely recall the audio of a metal cage being shaken violently, as if a small animal was agitated and restless. From what I remember, there was also a man's voice, casually repeating "Yeah," as if responding to the lively creature, which felt very routine.

MLLM output context:

<think>The user describes a majestic male lion in a transport cage on a truck with specific lighting. The audio description (violent shaking, small animal, man saying "Yeah") strongly resembles a viral meme involving a raccoon or possum, suggesting the user might be looking for a meme edit or a specific viral clip. I will search for the visual content of the lion transport and the specific audio context.</think>
<tool_call>{"name": "search_videos", "arguments": {"query": ["lion in transport cage truck majestic", "lion cage truck release video"]}}</tool_call>

Selected frame:



Ground truth text:

I want to find a video segment, the main subject of this frame is a large, majestic

adult male lion standing inside a metal transport cage mounted on the back of a truck. It has a thick, flowing mane, transitioning from golden yellow on the head to deep black around the neck, exuding strength and dignity. The lion stands tall, alert and solemn, gazing intently toward the upper right of the frame. The cage, featuring a black metal frame and dense silver wire mesh, is positioned above a lowered metal ramp and a heavy-duty white bumper with circular holes, creating a strong industrial and oppressive feel. Bright natural sunlight streams from the upper right, clearly highlighting the texture of the lion's facial muscles and mane.

Ground truth frame:

Model's prediction: False

Human's prediction: True

H Annotation Guidelines

The goal of this annotation task is to manually verify and refine model-generated shot descriptions to ensure that they accurately and faithfully reflect the corresponding video shots. Annotators are primarily responsible for identifying discrepancies between the generated descriptions and the actual video content, and correcting them when necessary.

1. Video–Description Alignment

Before making any edits, please carefully watch the corresponding video shot in full. Pay close attention to the main subjects, actions, scenes, and event progression, as well as relevant attributes such as color characteristics, visual style, audio content, and visual quality. Then, review the model-generated shot description and constraint descriptions to form an overall understanding of their alignment with the video.

2. Discrepancy Identification

Please assess whether the generated description contains any discrepancies with respect to the video content, considering the following aspects:

- Semantic consistency: Whether the description correctly captures the actions, events, and visual content shown in the video, without missing key elements or introducing nonexistent ones.
- Factual accuracy: Whether entities such as people, objects, scenes, and temporal order match the video.
- Constraint consistency: Whether constraints (e.g., temporal sequence, color, style, audio, and resolution) are consistent with what is actually observable in the video.

Any clear inconsistency in these aspects should be treated as a discrepancy.

3. Revision Principles

If discrepancies are found, please revise the description based strictly on the video content. Revisions should follow these principles:

- Accuracy first: All statements in the description must be directly supported by the video.
- Minimal necessary changes: Make only the modifications required to correct errors, and avoid unnecessary rewriting.
- Objectivity: Do not add speculative or inferred information that is not explicitly visible or audible in the video.
- Clarity and precision: When correcting errors, improve clarity and precision where appropriate.

4. Handling Correct Descriptions

If the generated description is fully consistent with the video and contains no factual or constraint-related errors, it may be retained without modification.

5. Annotation Objective

By following this verification and refinement process, the final descriptions should accurately correspond to their video shots, providing a high-quality, low-noise, and reproducible dataset for benchmark evaluation and subsequent research.

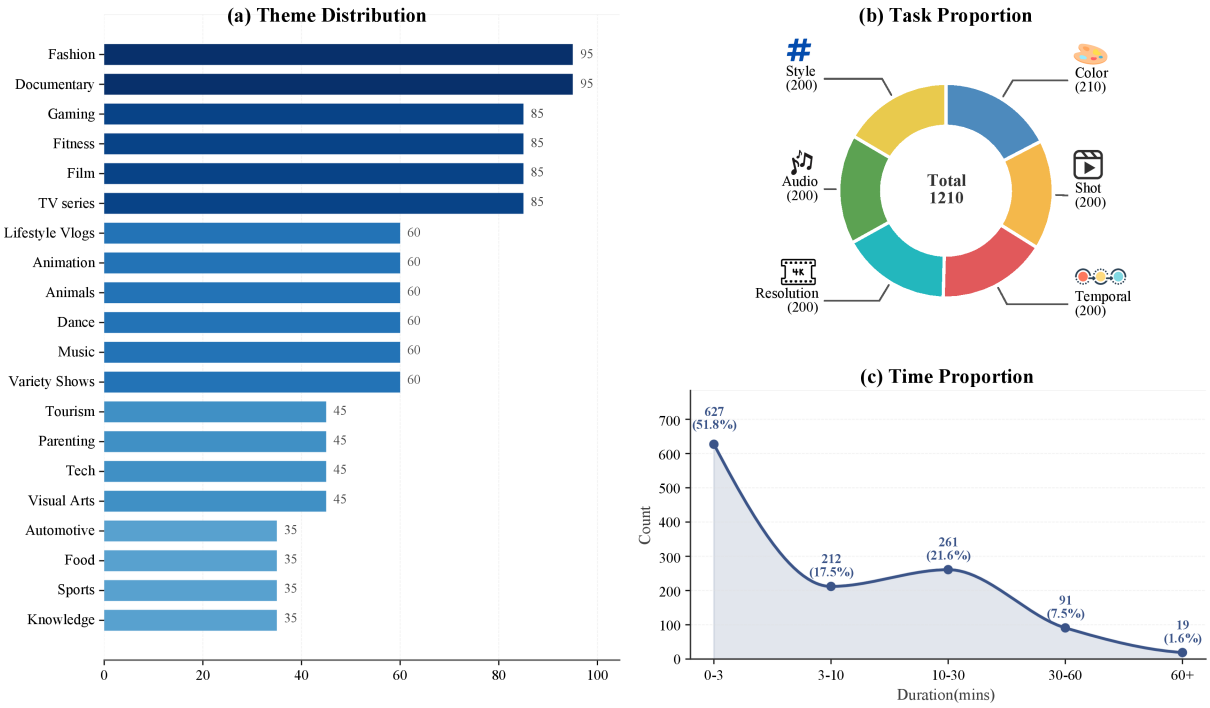


Figure 5: Theme distribution, task proportion, and time proportion of the dataset.

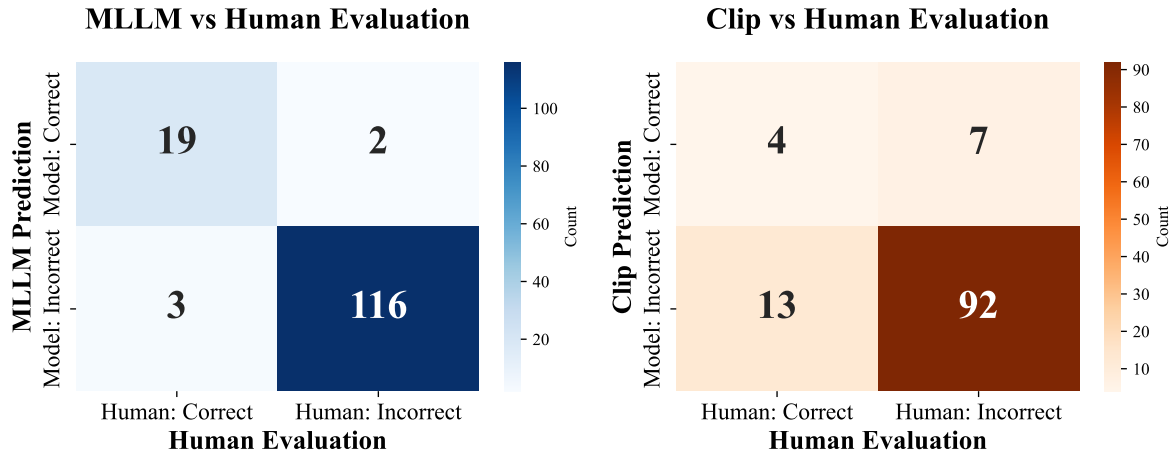


Figure 6: Confusion matrices between MLLM evaluation and human evaluation, and between CLIP evaluation and human evaluation

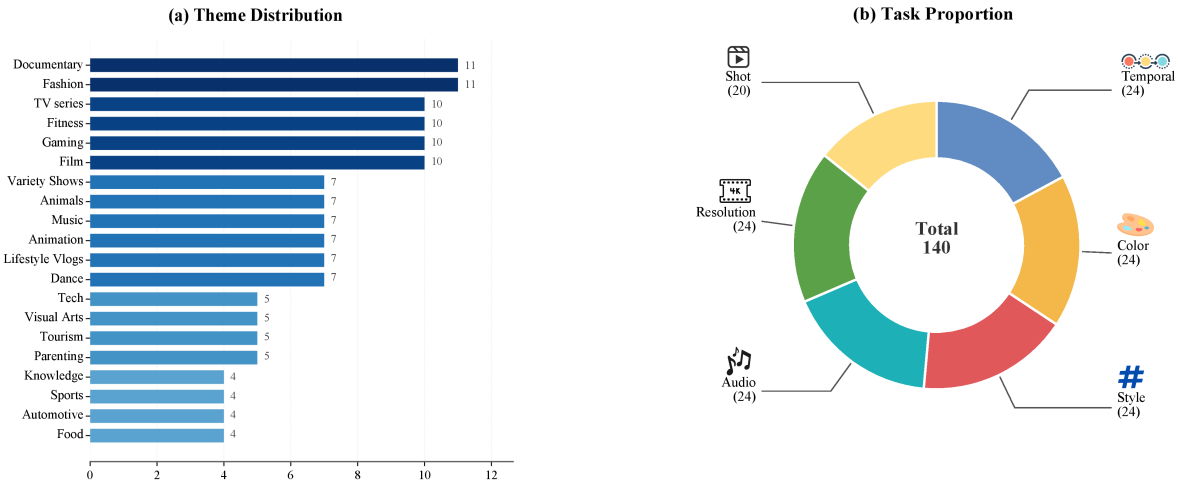


Figure 7: Theme distribution and task proportion of the dataset for further analysis.