Supplementary Materials for *MUVR: A Multi-Modal Untrimmed Video Retrieval Benchmark with Multi-Level Visual Correspondence*

Anonymous Author(s)
Affiliation
Address
email

In this supplementary material, we elaborate on the MLLMs prompting details in Section 1. We further illustrate the annotation instructions in Section 2. Then, some visualization examples are provided in Section 3. Limitations and social impact are introduced in Section 4.

1 MLLMs Prompting Details

The evaluation prompts for MLLMs are listed in Table 1 and 2. Although we attempted to maintain consistency across models, slight variations were necessary due to differing prompting requirements. The proprietary models (GPT-4o and Gemini-2.0-Flash) were accessed on April 25, 2025.

2 Annotation Instructions

The instructions provided to annotators are included below. We take the relationship annotation of the News partition as an example, while other partitions have different visual correspondences.

3 Visualization

Figure 1, 2, 3, 4 and 5 provide several relevant examples of different partitions from MUVR, with a text description of the query video and the tag of each video.

4 Limitations and Social Impact

Limitations. MUVR relies on human annotators to annotate videos with rich semantics. Despite strict guidance for annotators and multiple rounds of validation during the annotation process, there may still be minor annotation errors. Besides, MUVR focuses on visual and textual modalities, leaving out other potential modalities such as audio, which could further enrich the retrieval task. Despite these limitations, we believe MUVR offers a robust foundation for advancing research in video retrieval, and its design allows for future extensions to address these gaps.

Social Impact. The development of MUVR has potential positive implications for improving video search and recommendation systems, enhancing user experience on video-sharing platforms. By enabling more accurate and fine-grained retrieval, our work could facilitate better access to educational, informational, and entertainment content.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Format of the text prompts used by MLLMs for one-stage text query-only comparison. <Target Video>: format as 'Frame1: <image>\nFrame2: <image>\n...Frame6: <image>\n'.

Model	**Text Prompt**
InternVL2[1]	I will give you a text query and a video: [Query] and [Target]. Please determine whether any part of [Target] is slightly relevant to any part of [Query]. I will also provide [Tag] that [Target] (if relevant) must feature it.\n[Query]:\n{Text Description}\n[Target]:\n<Target Video>\n[Tag]:\n{Tag Prompt}\n[Output]:\n If slightly relevant, return Yes. If not, return No.
InternVL2.5[2]	I will give you a text query and a video: [Query] and [Target]. Please determine whether any part of [Target] is slightly relevant to any part of [Query]. I will also provide [Tag] that [Target] (if relevant) must feature it.\n[Query]:\n{Text Description}\n[Target]:\n<Target Video>\n[Tag]:\n{Tag Prompt}\n[Output]:\n If slightly relevant, return Yes. If not, return No.
MiniCPM-o 2.6[3]	Please determine whether any part of the video is slightly relevant to any part of [Query]. I will also provide [Tag] that the video (if relevant) must feature it. [Query]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
MiniCPM-V 2.6[4]	Please determine whether any part of the video is slightly relevant to any part of [Query]. I will also provide [Tag] that the video (if relevant) must feature it. [Query]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
LLaVA-NeXT-Video[5]	Please determine whether any part of the video is slightly relevant to any part of [Query]. I will also provide [Tag] that the video (if relevant) must feature it. [Query]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
LLaVA-OV[6]	Please determine whether any part of the video is slightly relevant to any part of [Query]. I will also provide [Tag] that the video (if relevant) must feature it. [Query]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
LLaVA-Video[7]	Please determine whether any part of the video is slightly relevant to any part of [Query]. I will also provide [Tag] that the video (if relevant) must feature it. [Query]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.

Table 2: Format of the text prompts used by MLLMs for one-stage multi-image comparison. <Query Video>/<Target Video>: format as 'Frame1: <image>\nFrame2: <image>\n...Frame6: <image>\n'. †: using additional mask prompt.

Model	**Text Prompt**
InternVL2[1]	I will give you a video query and a video target: [Query] and [Target]. Please determine whether any part of [Target] is slightly relevant to any part of [Query] or [Focus]. I will also provide [Tag] that [Target] (if relevant) must feature it.\n[Query]:\n<Query Video>\n[Target]:\n<Target Video>\n[Focus]\n{Text Description}\n[Tag]:\n{Tag Prompt}\n[Output]:\n If slightly relevant, return Yes. If not, return No.
InternVL2.5[2]	I will give you a video query and a video target: [Query] and [Target]. Please determine whether any part of [Target] is slightly relevant to any part of [Query] or [Focus]. I will also provide [Tag] that [Target] (if relevant) must feature it.\n[Query]:\n<Query Video>\n[Target]:\n<Target Video>\n[Focus]\n{Text Description}\n[Tag]:\n{Tag Prompt}\n[Output]:\n If slightly relevant, return Yes. If not, return No.
MiniCPM-o 2.6[3]	Please determine whether any part of <Target Video> is slightly relevant to any part of <Query Video> and [Focus]. I will also provide [Tag] that <Target Video> (if relevant) must feature it. [Focus]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
MiniCPM-V 2.6[4]	Please determine whether any part of <Target Video> is slightly relevant to any part of <Query Video> and [Focus]. I will also provide [Tag] that <Target Video> (if relevant) must feature it. [Focus]: {Text Description}\n[Tag]: {Tag Prompt}\n If slightly relevant, return Yes. If not, return No.
VideoRefer[8]	Here are two videos with same length. Is any part of the first video query slightly relevant to any part of the second video? {Text Description}\n If true and {Tag Prompt}, return Yes. Else, return No.
VideoRefer†[8]	Here are two videos with same length. Is any part of the first video query slightly relevant to any part of the second video? {Text Description}\n If true and {Tag Prompt}, return Yes. Else, return No.
Gemini-2.0-Flash[9]	Is any part of the first video query slightly relevant to any part of the second video? {Text Description}\n If true and {Tag Prompt}, return Yes. Else, return No.
GPT-4o[10]	Here are two videos with same length. Is any part of the first video query slightly relevant to any part of the second video? {Text Description}\n If true and {Tag Prompt}, return Yes. Else, return No.

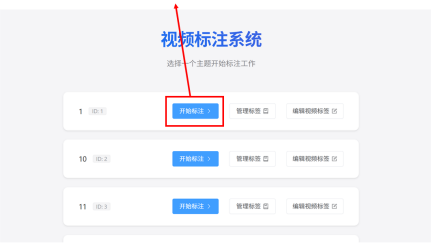
Annotation Guide

October 26 2025

Introduction

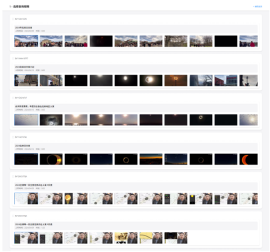
- **Your goal is to label whether a target video is relevant to a query video**
- **You can choose（"copy"，"event"，"copy and event"，"independent"）for the target video.**
- **You should label each video as accurately as possible.**

Click on this button to pick a topic.

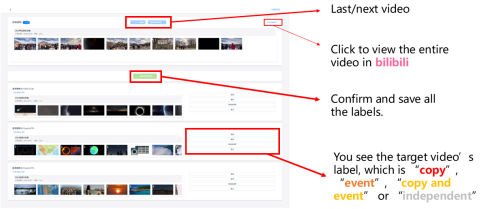


In this page, all the videos of the topic are shown with some **keyframes**.

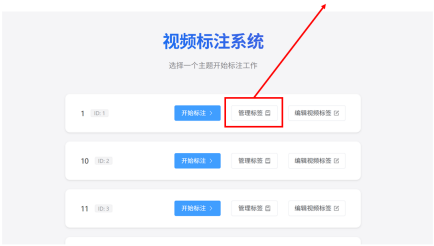
You should choose **three videos** that have **strong correlation** to the topic as the query videos.



Choose the relationship between the **target video** above and the **three query videos** below respectively

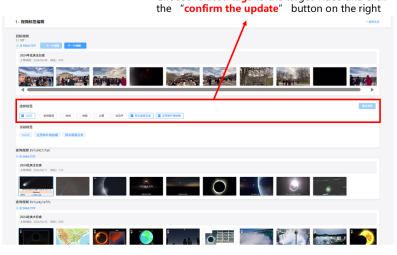


Click to **set tags of the videos**



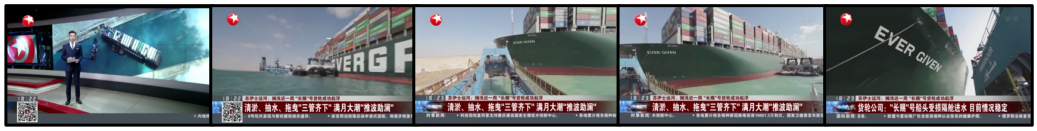
Click here to **tag each video separately**

Choose **related tags** to the target video and click the **"confirm the update"** button on the right



Text Description: Please pay attention to the Evergiven photographed from a distance in the query. There are a large number of containers on the ship and the hull has the words EVERGIVEN.

Tag: Self-media news reports



Tag: TV news reports

Tag: Aerial photography of the entire ship

Figure 1: Visualization of three relevant videos on the News partition.

Text Description: Please pay attention to the towering tower of Notre Dame Cathedral in Query, the middle is connected by exquisite stone decorations, and the sculptures on the exterior wall are exquisitely crafted.

Tag: Double tower structure

Tag: Aerial shots; Partial close-up

Tag: Aerial shots; Notre Dame model; Partial close-up; Double tower structure

Figure 2: Visualization of three relevant videos on the Region partition. The third video comes from a computer game and brings more challenges.

Text Description: Please pay attention to the phone in the query that can be folded, the body is red, the edges are decorated with gold lines, and the camera area is black.

Tag: Red shell; Folding screen

Tag: Red shell; Folding screen; Blogger introduction screen

Tag: Red shell

Figure 3: Visualization of three relevant videos on the Instance partition. The different forms of mobile phones and their small proportion on the screen pose challenges.

Text Description: Please pay attention to the folding of the wrist in the query, and the palms are used to build geometric figures. There seems to be a 3×3 grid in front of you. The palms are used to build the figures based on the lines of these grids.

Tag: Arm formation 90°

Tag: Arm formation 90°

Tag: Fold the wrist and build geometric shapes with palms

Figure 4: Visualization of three relevant videos on the Dance partition. The interference of background, characters, and the number of people poses a huge challenge to action-level retrieval.

Text Description: Please pay attention to the solid background and white text in query. The text takes up a large area of the picture. The cartoon character walks with his legs raised, holding a musical instrument in his hands, playing the flute, and the character's forehead is exposed.

Tag: Cartoon character; Little girl wearing white socks, white shirt and red dress carrying a red backpack

Tag: Little girl wearing white socks, white shirt and red dress carrying a red backpack

Tag: Cartoon character

Figure 5: Visualization of three relevant videos on the Others partition. This type of video is created based on common popular elements and video styles, with rich semantic information.

References

[1] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi *et al.*, "Internvideo2: Scaling foundation models for multimodal video understanding," in *European Conference on Computer Vision*. Springer, 2024, pp. 396–416.

[2] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[3] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao *et al.*, "Minicpm: Unveiling the potential of small language models with scalable training strategies," *arXiv preprint arXiv:2404.06395*, 2024.

[4] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[5] Y. Zhang, B. Li, H. Liu, Y. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, "Llava-next: A strong zero-shot video understanding model," 2024.

[6] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[7] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[8] Y. Yuan, H. Zhang, W. Li, Z. Cheng, B. Zhang, L. Li, X. Li, D. Zhao, W. Zhang, Y. Zhuang *et al.*, "Videorefer suite: Advancing spatial-temporal object understanding with video llm," *arXiv preprint arXiv:2501.00599*, 2024.

[9] G. Team, "Gemini: A family of highly capable multimodal models," 2025.

[10] OpenAI, "Gpt-4o," 2024, accessed: 2024-11-15.