Supplementary Material of Compositional Transformation Reasoning for Composed Video Retrieval

1. Prompt Templates

1.1. Target Video Narration Generation

As shown in Tables 1– 7, we present in detail the complete process of generating target video narrations using a Multimodal Large Language Model (MLLM). Unlike existing description-generation approaches, our method incorporates not only the textual modalities of the reference video narration and the modification text but also the visual modality of the reference video. Specifically, we first independently generate a scene-centric global description, an action-centric local description, and an entity-centric local description. These three descriptions are then semantically transformed according to the given modification text. Finally, the modified scene, action, and entity descriptions are fused to produce the final target video narration. This narration holistically captures the semantics inferred by the MLLM from the reference video, its narration, and the modification text, thereby improving the recall of text-to-video retrieval for target videos. Tables 8- 11 provide concrete examples of the target video narrations generated using the above method.

p_{Scene}
1. Describe only the scene of the video: room type, background objects, lighting environment, spatial location, functional cues. 2. Do not describe the entity or action. 3. Do not speculate; describe only what you see in the video. 4. Write the description as a single paragraph (one paragraph). **Input**: "q" **Output**: "C_q"

Table 1. Reference video Scene-centric generation prompt.

p_{Action}
1. Describe only the object actions/motions and temporal dynamics in the video: behaviors, interactions, gestures, movement trajectories. 2. Do not describe the entity or scene/background. 3. Do not speculate; describe only what you see in the video. 4. Write the description as a single paragraph (one paragraph). **Input**: "q" **Output**: "A_q"

Table 2. Reference video Action-centric generation prompt.

p_{Entity}

1. Describe only the entity of the video: category, attributes, appearance, parts, position.
2. Do not mention actions or scenes. Be objective and detailed.
3. Do not speculate; describe only what you see in the video.
4. Write the description as a single paragraph (one paragraph).

Input: "q"
Output: "E_q"

Table 3. Reference video Entity-centric generation prompt.

$p_{SChange}$

Goal: In the Scene description C_q, LOCATE the sentences/phrases that semantically correspond to n_q; DELETE those matched parts; then INSERT t in their place.

1. Alignment: Map n_q to the most semantically corresponding content in C_q.
2. Deletion+Insertion: Remove the matched spans from C_q, and insert t exactly (do not paraphrase t).
3. Preservation: Keep all other content in C_q unchanged; do NOT invent facts beyond C_q/n_q/t.
4. Form: Output the update Scene-centric description as one paragraph.

Input: C_q, n_q, t
Output: $\hat{C}_t$

Table 4. Edit Scene-centric Narration prompt.

$p_{AChange}$

Goal: Update the Action-centric description A_q by first identifying and replacing content tied to n_q with t.

1. Alignment: Try to align n_q to semantically corresponding sentences/phrases in A_q. If found, delete the matched spans and insert t exactly in their place.
2. Deletion+Insertion: If A_q contains no actions consistent with t, then insert t as the action content and delete all other actions.
3. Preservation: Keep all other content in A_q unchanged; do not invent facts beyond A_q/n_q/t.
4. Form: Output the updated Action-centric description as one paragraph.

Input: A_q, n_q, t
Output: $\hat{A}_t$

Table 5. Edit Action-centric Narration prompt.

$p_{EChange}$

Goal: Update the Entity-centric description E_q by first identifying and replacing content tied to n_q with t.

1. Alignment: Try to align n_q to semantically corresponding sentences/phrases in E_q. If found, delete the matched spans and insert t exactly in their place.
2. Deletion+Insertion: If E_q contains no entity consistent with t, then insert t as the entity content.
3. Preservation: Keep all other content in E_q unchanged; do not invent facts beyond E_q/n_q/t.
4. Form: Output the updated Entity-centric description as one paragraph.

Input: E_q, n_q, t
Output: $\hat{E}_t$

Table 6. Edit Entity-centric Narration prompt.

p_{Fusion}

Generate the target video narration according to the following requirements.
You are given three descriptions from the SAME video: a scene-centric description $\hat{C}_t$, an action-centric description $\hat{A}_t$, and an entity-centric description $\hat{E}_t$.
Task: Produce one coherent, natural, and non-redundant narration that integrates $\hat{C}_t$, $\hat{A}_t$ and $\hat{E}_t$.
1. Keep salient details from $\hat{C}_t$, $\hat{A}_t$ and $\hat{E}_t$.
2. Write as a single paragraph, objective tone, third person, concise but information-dense.
Input: $\hat{C}_t$, $\hat{A}_t$, $\hat{E}_t$
Output: $\hat{s}_t$

Table 7. Fusion Scene-Action-Entity Narration prompt.

$1 - Example_{TargetVideoNarrationGeneration}$

$\boldsymbol{n_q}$: Takes the sieve.
$\boldsymbol{t}$: Shake it.
$\boldsymbol{C_q}$: The video is set in a cozy kitchen environment with a warm and inviting atmosphere. The camera viewpoint is from above, providing a bird's-eye view of the scene. The kitchen features a gas stove with a pot on it, a wooden countertop holding various items, a blue floor below the countertop, a dish rack filled with clean dishes to the left, and a wooden shelf holding additional kitchenware to the right. The lighting appears to be warm and natural, suggesting it might be daytime. The overall setting gives off a homely and functional vibe, typical of a lived-in kitchen space.
$\boldsymbol{A_q}$: The person is cooking at the stove, stirring a pot with one hand while holding a glass lid in the other. They then move to the counter, placing the lid down and reaching for an orange bowl. The person appears to be preparing ingredients or utensils, possibly for the next step in their cooking process.
$\boldsymbol{E_q}$: Various items are placed on the wooden countertop, including a smartphone displaying an app, a metal strainer, and some kitchen utensils. A person's bare feet can be seen on the blue floor.
$\hat{\boldsymbol{C}}_{\boldsymbol{t}}$: The video is set in a cozy kitchen environment with a warm and inviting atmosphere. The camera viewpoint is from above, providing a bird's-eye view of the scene. The kitchen features a gas stove with a pot on it, a wooden countertop holding various items such as a smartphone displaying an app, a metal strainer, and some kitchen utensils. Below the countertop, there is a blue floor where a person's bare feet can be seen. To the left, there is a dish rack filled with clean dishes, and to the right, a wooden shelf holds additional kitchenware. The lighting appears to be warm and natural, suggesting it might be daytime. The overall setting gives off a homely and functional vibe, typical of a lived-in kitchen space.
$\hat{\boldsymbol{A}}_{\boldsymbol{t}}$: The person is cooking at the stove, stirring a pot with one hand while holding a glass lid in the other. They then move to the counter, placing the lid down and shaking the sieve. The person appears to be preparing ingredients or utensils, possibly for the next step in their cooking process.
$\hat{\boldsymbol{E}}_{\boldsymbol{t}}$: The person holds the sieve while preparing ingredients at the counter. A smartphone displaying an app, a metal strainer, and several kitchen utensils remain on the wooden countertop. The person's bare feet are visible on the blue floor beneath the counter.
$\hat{\boldsymbol{s}}_{\boldsymbol{t}}$: The video shows a cozy kitchen with a warm and inviting atmosphere, viewed from above. The person is cooking at the stove, stirring a pot with one hand while holding a glass lid in the other. They then move to the counter, place the lid down, and shake the sieve.

Table 8. Example-1: Target Video Narration Generation.

$2 - Example_{TargetVideoNarrationGeneration}$

$\boldsymbol{n_q}$: Digital network, white lines and dots.
$\boldsymbol{t}$: replace the white lines and dots with green.
$\boldsymbol{C_q}$: The video features a black background with an evolving network of white lines and dots that form a complex digital grid-like structure. The environment is abstract, resembling a technological or scientific visualization. There are no discernible objects, locations, or specific settings beyond this digital representation. The lighting is consistent, with the white elements standing out sharply against the dark backdrop. The camera viewpoint remains static throughout the sequence, focusing solely on the evolving network.
$\boldsymbol{A_q}$: The video begins with a black screen, which then transitions to an abstract display of white lines and dots forming a complex network structure. The lines and dots gradually expand and intertwine, creating a dynamic and evolving pattern that resembles a digital grid or circuit board. As the sequence progresses, the complexity of the network increases, with more intricate connections forming between the dots. The overall effect is one of growth and development, as if the network is expanding and becoming more interconnected over time. The video concludes with the network still in motion, showcasing its intricate and ever-evolving nature.
$\boldsymbol{E_q}$: There are no discernible objects, locations, or specific settings beyond this digital representation.
$\hat{\boldsymbol{C}}_{\boldsymbol{t}}$: The video features a black background with an evolving network of green lines and dots that form a complex digital grid-like structure. The environment is abstract, resembling a technological or scientific visualization. There are no discernible objects, locations, or specific settings beyond this digital representation. The lighting is consistent, with the green elements standing out sharply against the dark backdrop. The camera viewpoint remains static throughout the sequence, focusing solely on the evolving network.
$\hat{\boldsymbol{A}}_{\boldsymbol{t}}$: The video begins with a black screen, which then transitions to an abstract display of green lines and dots forming a complex network structure. The lines and dots gradually expand and intertwine, creating a dynamic and evolving pattern that resembles a digital grid or circuit board. As the sequence progresses, the complexity of the network increases, with more intricate connections forming between the dots. The overall effect is one of growth and development, as if the network is expanding and becoming more interconnected over time. The video concludes with the network still in motion, showcasing its intricate and ever-evolving nature.
$\hat{\boldsymbol{E}}_{\boldsymbol{t}}$: There are no discernible physical objects or identifiable entities in the scene; the only visible elements are the green lines and dots that constitute the abstract digital network.
$\hat{\boldsymbol{s}}_{\boldsymbol{t}}$: The video features a black background with an evolving network of green lines and dots that form a complex digital grid-like structure. The environment is abstract, resembling a technological or scientific visualization. The lighting is consistent, with the green elements standing out sharply against the dark backdrop. The camera viewpoint remains static throughout the sequence, focusing solely on the evolving network.

Table 9. Example-2: Target Video Narration Generation.

$3 - Example_{TargetVideoNarrationGeneration}$

$\boldsymbol{n_q}$: Shuffles a card.
$\boldsymbol{t}$: Place it on the table.
$\boldsymbol{C_q}$: The scene is set indoors on a wooden table with a light brown finish, featuring visible wood grain patterns. The environment appears to be a casual setting, possibly a home or a small gathering space. In the upper right corner, part of a woven basket is visible, adding a touch of rustic decor. The lighting is natural, likely from a nearby window, casting soft shadows across the table surface. The camera viewpoint is from above, providing an overhead perspective of the table and the hands interacting with the objects.
$\boldsymbol{A_q}$: The person in the foreground is shuffling a deck of cards with both hands, moving them back and forth to mix the cards thoroughly. The other individual's hands are visible on the table, initially clenched into fists but then opening slightly as if preparing to receive the cards. The shuffling continues smoothly, indicating a deliberate and practiced motion. The interaction suggests a card game is about to begin.
$\boldsymbol{E_q}$: On the left side of the frame, there is a red box with yellow and white text, partially visible, suggesting it might contain game pieces or cards. A yellow and black object, possibly a game piece container, is also present near the red box.
$\hat{\boldsymbol{C}}_{\boldsymbol{t}}$: The scene is set indoors on a wooden table with a light brown finish, featuring visible wood grain patterns. The environment appears to be a casual setting, possibly a home or a small gathering space. On the left side of the frame, there is a red box with yellow and white text, partially visible, suggesting it might contain game pieces or cards. A yellow and black object, possibly a game piece container, is also present near the red box. In the upper right corner, part of a woven basket is visible, adding a touch of rustic decor. The lighting is natural, likely from a nearby window, casting soft shadows across the table surface. The camera viewpoint is from above, providing an overhead perspective of the table and the hands interacting with the objects. The person in the foreground is shuffling a deck of cards with both hands, moving them back and forth to mix the cards thoroughly. The other individual's hands are visible on the table, initially clenched into fists but then opening slightly as if preparing to receive the cards. The shuffling continues smoothly, indicating a deliberate and practiced motion. The interaction suggests a card game is about to begin. The person shuffles a card and places it on the table.
$\hat{\boldsymbol{A}}_{\boldsymbol{t}}$: The person in the foreground is placing a card on the table with one hand, as indicated by the modification instruction. The other individual's hands remain visible on the table, initially clenched into fists but then opening slightly as if preparing to receive the card. The action suggests a card game is about to begin.
$\hat{\boldsymbol{E}}_{\boldsymbol{t}}$: On the left side of the frame, there is a red box with yellow and white text, partially visible, suggesting it might contain game pieces or cards. A yellow and black object, possibly a game piece container, is also present near the red box. The card that was previously being shuffled is now placed on the table.
$\hat{\boldsymbol{s}}_{\boldsymbol{t}}$: The scene is set indoors on a wooden table with a light brown finish, featuring visible wood grain patterns. The environment appears to be a casual setting, possibly a home or a small gathering space. On the left side of the frame, there is a red box with yellow and white text, partially visible, suggesting it might contain game pieces or cards. A yellow and black object, possibly a game piece container, is also present near the red box. In the upper right corner, part of a woven basket is visible, adding a touch of rustic decor. The lighting is natural, likely from a nearby window, casting soft shadows across the table surface. The person in the foreground is placing a card on the table with one hand, as indicated by the modification instruction, while the other individual's hands remain visible on the table, initially clenched into fists but then opening slightly as if preparing to receive the card. The interaction suggests a card game is about to begin.

Table 10. Example-3: Target Video Narration Generation.

$4 - Example_{TargetVideoNarrationGeneration}$
$\boldsymbol{n_q}$: Cows on pasture. $\boldsymbol{t}$: Replace cow with goat. $\boldsymbol{C_q}$: The video depicts a serene rural setting with a lush green field that appears expansive, with varying shades of grass suggesting a healthy pasture. The background consists of more open grassland, indicating a vast agricultural area. The lighting suggests it is daytime, likely late afternoon given the softness of the light and the long shadows. There are no visible structures or other animals in the immediate vicinity, emphasizing the peaceful and isolated nature of the scene. A red post is visible on the right side of the frame, possibly part of a fence or boundary marker for the pasture. $\boldsymbol{A_q}$: The video captures a serene pastoral scene where a group of cows is grazing in a lush green field. The cows, varying in color from white to brown and black with some having distinctive markings on their faces, are scattered across the frame, each engaged in feeding on the abundant grass. Their movements are slow and deliberate, reflecting a calm and unhurried pace typical of grazing animals. As they graze, some cows occasionally lift their heads, perhaps to look around or to interact with one another, while others continue to focus on their meal. The overall atmosphere conveys a sense of tranquility and natural rhythm, with the cows' actions subtly shifting throughout the frames, showcasing the peaceful coexistence of these animals in their environment. $\boldsymbol{E_q}$: In the foreground, there are several cows of different colors including white, brown, black, and mixed hues. Some cows are actively eating while others appear to be standing still or moving slightly. $\hat{\boldsymbol{C}}_{\boldsymbol{t}}$: The video depicts a serene rural setting with goats grazing in a lush green field. The field is expansive, with varying shades of green grass, suggesting a healthy pasture. In the foreground, there are several goats of different colors including white, brown, black, and a mix of these hues. Some goats are actively eating while others appear to be standing still or moving slightly. The background consists of more open grassland, indicating a vast agricultural area. The lighting suggests it is daytime, likely late afternoon given the softness of the light and long shadows cast by the goats. There are no visible structures or other animals in the immediate vicinity, emphasizing the peaceful and isolated nature of the scene. A red post is visible on the right side of the frame, possibly part of a fence or boundary marker for the pasture. $\hat{\boldsymbol{A}}_{\boldsymbol{t}}$: The video captures a serene pastoral scene where a group of goats is grazing in a lush green field. The goats, varying in color from white to brown and black with some having distinctive markings on their faces, are scattered across the frame, each engaged in feeding on the abundant grass. Their movements are slow and deliberate, reflecting a calm and unhurried pace typical of grazing animals. As they graze, some goats occasionally lift their heads, perhaps to look around or to interact with one another, while others continue to focus on their meal. The overall atmosphere conveys a sense of tranquility and natural rhythm, with the goats' actions subtly shifting throughout the frames, showcasing the peaceful coexistence of these animals in their environment. $\hat{\boldsymbol{E}}_{\boldsymbol{t}}$: In the foreground, there are several goats of different colors including white, brown, black, and mixed hues. Some goats are actively eating while others appear to be standing still or moving slightly. $\hat{\boldsymbol{s}}_{\boldsymbol{t}}$: The video depicts a serene rural setting with goats grazing in a lush green field. The goats, varying in color from white to brown and black with some having distinctive markings on their faces, are scattered across the frame, each engaged in feeding on the abundant grass. Their movements are slow and deliberate, reflecting a calm and unhurried pace typical of grazing animals. The overall atmosphere conveys a sense of tranquility and natural rhythm, with the goats' actions subtly shifting throughout the frames, showcasing the peaceful coexistence of these animals in their environment.

Table 11. Example-4: Target Video Narration Generation.

1.2. Candidate Video Reasoning with MLLM

As shown in Table 12, to evaluate the reasoning capability of the MLLM in candidate video ranking, we adopt the Relative-Comparison setting. In this setting, the model determines which candidate video better matches the semantic modification described by the instruction, while also providing the options tie (both candidates match equally well) and none (neither video is relevant) to allow more flexible judgment. This setup takes the reference video, its narration, and the modification text as inputs, and outputs the identifier of the candidate video that best reflects the intended semantic change.

Figure 1 presents examples of composed video retrieval results. As shown, existing supervised and training-free methods fail to achieve correct semantic transformations simultaneously across the entity, action, and scene dimensions. In con-

trast, our proposed method produces correct transformations on all three dimensions, demonstrating the effectiveness of our approach.

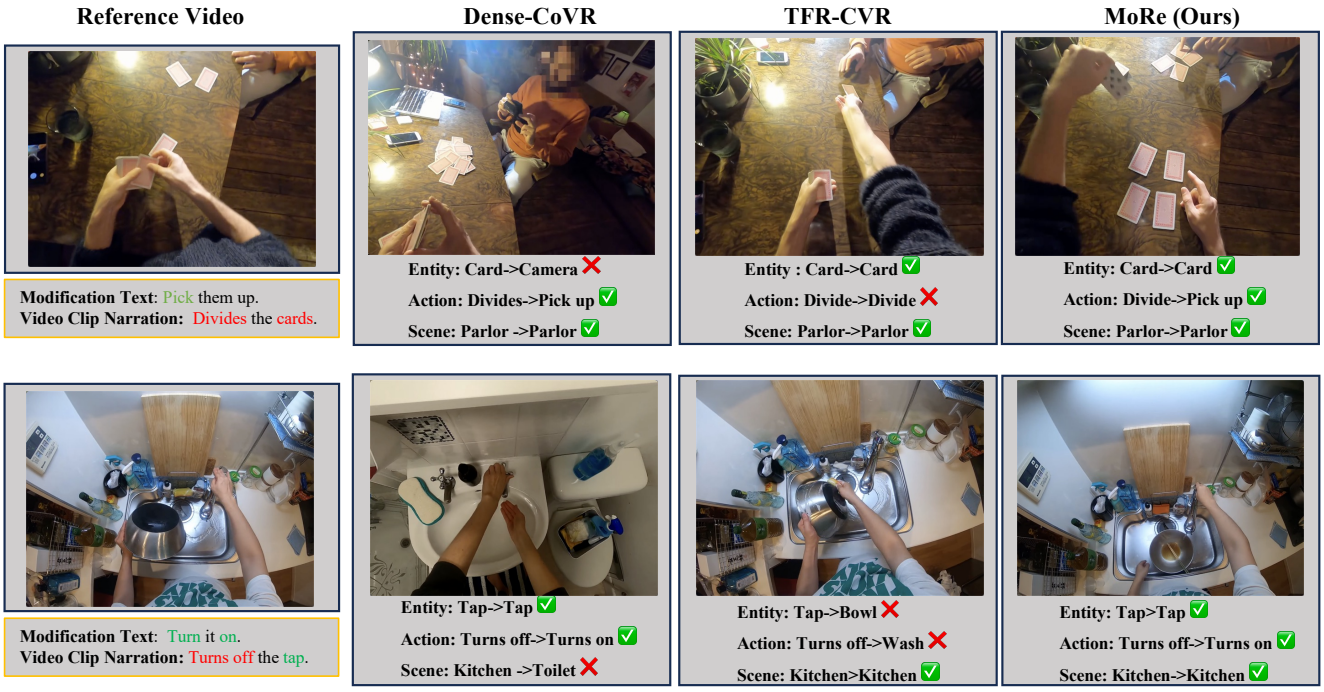

Figure 1. Examples of composed video retrieval.

p_{cmp}
You are a composed video retrieval task expert. Given the following reference video q, reference video narration n_q and modification text t describing a desired change or target, which of the candidate videos best matches the intended change? This task consists of the following steps: 1. Identify the entity, action, and scene from the reference video and its narration. 2. Determine which dimensions are changed by the modification text. 3. Update the changed dimension. 4. Compare both candidate videos with the updated entity, action, and scene, then select the candidate video that best matches the target video. If both candidates are equally matched, choose the option 'tie'. If neither is related to the reference, choose the option 'Both videos are not related to the reference video and text' Please provide only the single option letter (e.g., A, B, C, D, etc.). A. candidate video v_i B. candidate video v_j C. tie D. Both videos are not related to the reference video and text **Input**: q, n_q, t, v_i, v_j **Output**: $o_{i,j}$

Table 12. Video ranking based on pairwise comparisons under the relative comparison setting of the MLLM.

1.3. Result of Different Text to Video retrieval

Figure 2 presents examples of target video retrieval results using different textual inputs. As shown, the texts produced by existing methods (Modification Text [2, 3] fail to retrieve videos that are simultaneously correct across the entity, action, and scene dimensions. In contrast, our proposed method successfully retrieves videos that are correct on all three dimensions. This demonstrates that our target video narration generation approach effectively captures richer contextual information, thereby improving the accuracy of text-based video retrieval.

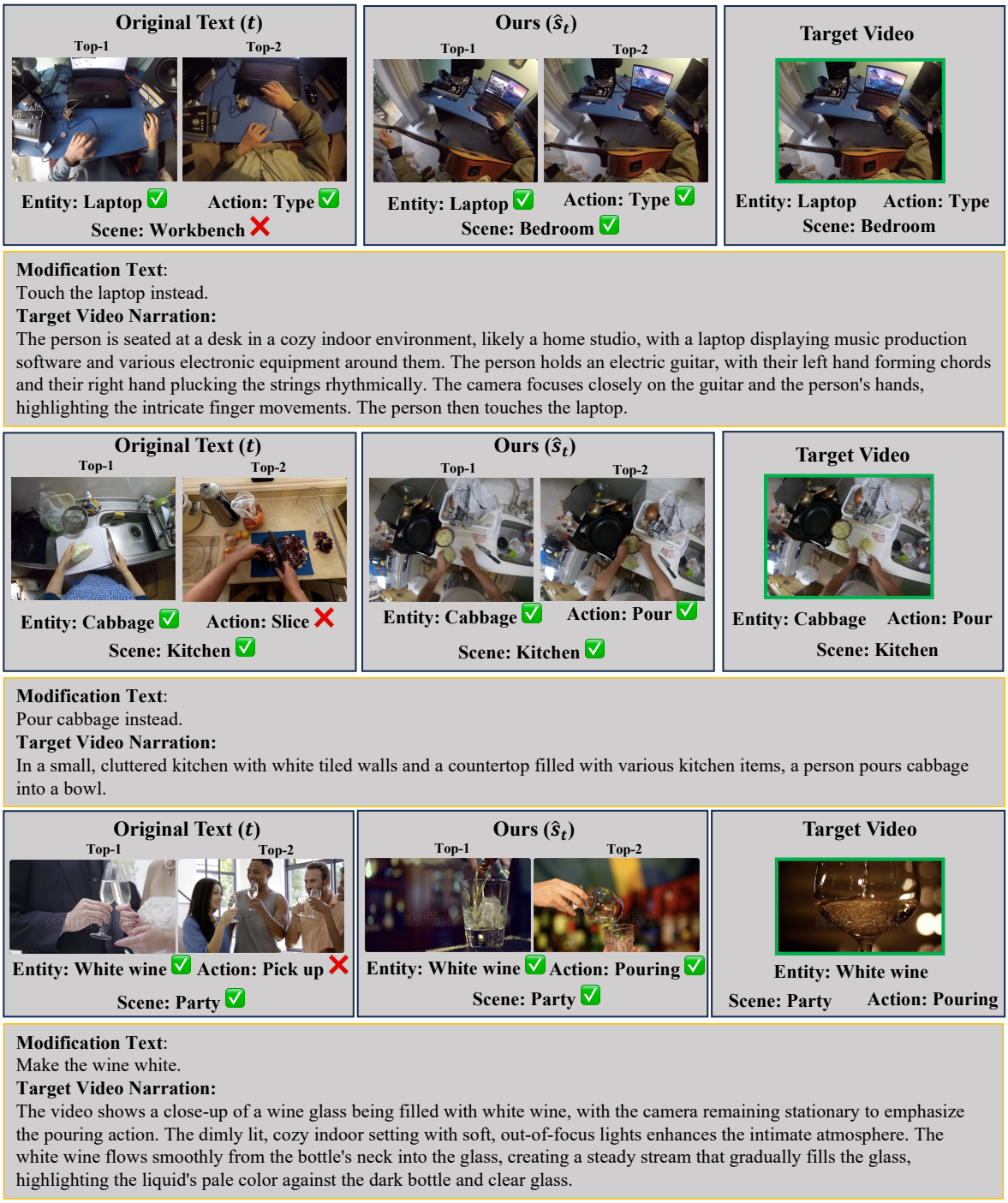

Figure 2. Examples of Text-to-Video retrieval.

2. Dynamic Weighting VS ROVCS Module

Table 13. Dynamic Weighting VS ROVCS Module.

	R@1	R@5	R@10
MoRe(Dynamic Weighting)	15.3	44.31	51.54
MoRe(ROVCS)	**20.4**	**53.9**	**72.1**

As shown in Table 13, we replace the ROVCS module with an MLLM-based dynamic objective weighting module, and the resulting performance drops to 15.3, compared with 20.4 achieved by ROVCS. Further analysis of the candidate video set recall shows that the dynamic objective weighting module attains only 56.2, whereas ROVCS achieves a substantially higher recall of 74.4.

3. Ablation Studies on Fine-grained Decomposition

Table 14. Impact of Entity, Action and Scene.

	R@1	R@5	R@10
Only Entity	17.6	47.1	64.5
Only Action	17.9	50.3	65.3
Only Scene	10.8	35.8	57.6
Entity + Action	18.5	48.1	65.8
Entity + Scene	19.1	**54.2**	68.4
Action + Scene	18.2	45.1	63.4
Entity + Action + Scene	**20.4**	53.9	**72.1**

As shown in Table 14, we conduct leave-one-out experiments on the semantic components on the EgoCVR dataset. The R@1 values for entity-only, action-only, and scene-only are 17.6, 17.9, and 10.8, respectively. When combining two of them, namely entity+action, entity+scene, and action+scene, the R@1 rises to 18.5, 19.1, and 18.2, respectively. When using all three together, the performance further improves to 20.35. The results indicate that the entity and action dimensions are more effective than the scene dimension when used alone. However, when two dimensions are combined, the performance differences become relatively small. Combining all three yields the best result.

4. Result on different MLLM

Table 15. Performance comparison on different MLLM

MLLM	EgoCVR			WebVid-CoVR		
	R@1	R@5	R@10	R@1	R@5	R@10
Qwen2.5VL [1]	**20.4**	**53.9**	**72.1**	**63.0**	**83.4**	87.6
mPLUG-Owl3 [4]	18.3	47.5	65.2	45.1	74.3	**89.3**

We evaluate the impact of different MLLM on the proposed method for the CoVR task. Specifically, we instantiate the re-ranking stage with Qwen2.5-VL [1] and mPLUG-Owl3 [4], respectively. As shown in Table 15, although mPLUG-Owl3 yields lower precision than Qwen2.5-VL, its recall still surpasses all existing baselines. Moreover, when Qwen2.5-VL is used as the MLLM for re-ranking, our method outperforms all baselines on overall metrics, thereby validating the effectiveness of the proposed approach.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 9

[2] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 8

[3] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 8

[4] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 9