

Beyond Simple Edits: Composed Video Retrieval with Dense Modifications

(Paper ID : 11454)

Supplementary Material

A. Supplementary Materials

In this appendix, we provide a detailed discussion of the related work on both Composed Image Retrieval (CoIR) and Composed Video Retrieval (CoVR). CoVR, as an extension of CoIR, brings the challenge of video temporal dynamics and detailed modifications into the retrieval space. We also present dataset statistics for our proposed fine-grained composed video retrieval dataset, highlighting the significant increase in the number of words in modification-text and the length of descriptions compared to the WebVid-CoVR dataset. In particular, our dataset provides substantially more detailed modification texts, averaging around 80 words per description, as compared to the sparse and shorter descriptions in WebVid-CoVR. This rich data allows for better handling of fine-grained video modifications, significantly improving retrieval performance.

B. More on Related Work

Composed image retrieval has evolved significantly with various approaches leveraging visual and textual inputs for accurate retrieval. Early methods like TIRG [Vo et al., 2019] employed hybrid embeddings but struggled with fine-grained details. Recent models, such as CLIP [Radford et al., 2021] and ARTEMIS [Delmas et al., 2022], improved retrieval by aligning images and text, though they often required large-scale pretraining. Cross-attention-based methods like CompoDiff [Gu et al., 2023] have shown promising results by better integrating visual and textual information. Our approach builds on these advancements by introducing dense modification texts, offering superior performance over state-of-the-art methods like BLIP [Li et al., 2023] and Thawakar et al. [2024], particularly in fine-grained and zero-shot retrieval tasks.

Several notable datasets have been developed to advance CoVR. EgoCVR [2], one of the first benchmarks, emphasizes fine-grained action retrieval from egocentric videos, requiring temporal reasoning for accurate retrieval. WebVid-CoVR [5] offers a large-scale dataset created automatically by mining video pairs with similar captions and generating modification texts using large language models, resulting in 1.6 million triplets. This dataset enables scalable, high-quality annotations for general-purpose CoVR tasks. Recently, [4] proposes detailed language descriptions for source and target videos, improving context preservation and query-specific alignment. This work expands on WebVid-CoVR with detailed captions, further enhancing

multi-modal embedding alignment and retrieval accuracy. In contrast, this work proposes a fine-grained CoVR dataset that offers dense and detailed modification texts, enabling models to capture subtle visual and temporal changes between videos. Unlike previous datasets, which often rely on sparse annotations, our approach provides comprehensive and context-rich data, making it a useful resource for training more accurate and robust retrieval models. The scalability and depth of our dataset push the boundaries of composed video retrieval, facilitating more precise video search across a wide range of real-world applications, from video editing to surveillance.

C. Prompt used for Detailed Video Description Generation

For generating detailed video descriptions, we utilized the WebVid-CoVR dataset, comprising diverse categories such as nature, lifestyle, and professional activities. The dataset's videos, averaging around 16.8 seconds in length, offer substantial diversity in terms of visual content and actions. To effectively create captions for such videos, we employed Gemini-Pro [3], a leading-edge video captioning model particularly suited for handling longer and more intricate videos. Unlike other models that often falter when processing extensive sequences, Gemini-Pro excels in maintaining a coherent narrative over time, making it ideal for this task. Its advanced temporal context-handling abilities ensure that even videos with numerous transitions and complex actions are described with high accuracy and detail.

The video descriptions were generated using the prompt shown in Figure 1. This prompt was designed to extract comprehensive descriptions, ensuring the inclusion of all relevant details within the video. To ensure the quality and accuracy of the generated captions, we implemented a hallucination check using the BLIP model, which computes the cosine similarity between the input video and its corresponding caption. Video captions with a similarity score below 0.4 were considered inadequate and recomputed. This additional validation step ensures that the captions are both highly aligned with the video content and free from irrelevant information, resulting in a more robust and reliable dataset.

D. Prompt used for detailed Modification Text Generation

In fine-grained CoVR, modification text plays a key role in distinguishing two similar videos by clearly describing their differences, such as changes in actions, objects, or scenes. These texts focus on subtle variations to improve video retrieval based on fine details. To achieve this, we use GPT-4o [1] to generate precise modification texts. The model is used with the help of in context learning using default triplets from WebVid, which include original video captions and their corresponding short modification texts, ensuring accuracy in generated detailed modification texts. The prompt shown in figure 2 was utilized to generate detailed modification texts. By leveraging this detailed prompt and using WebVid-CoVR triplets as a reference, GPT-4o was able to produce high-quality, contextually rich modification texts.

E. Quality Control Protocol for Human Verification of the generated triplets

The manual human verification process involves trained annotators who systematically validate and refine the generated modification texts. Annotators are presented with side-by-side input and target videos, along with their corresponding modification text, and are responsible for assessing the correctness and completeness of the descriptions. To enhance inter-annotator reliability, we employ a consensus-based approach, where a subset of the dataset undergoes multiple rounds of validation. In order to assess the quality of the generated data and increase reliability, we provide a human-verified evaluation set, where each triplet is manually checked. Specifically, the overall verification pipeline and qualitative properties of the generated data are depicted in Figure 3. To ensure the correctness of both automated and manual modifications, we employ the following multi-check verification protocol:

- **Side-by-Side Comparison:** Present the input and target videos alongside their corresponding modification text to the reviewers. Ensure that the modifications mentioned accurately reflect the differences between the two videos.
- **Contextual Consistency Check:** Verify that the modification text addresses key changes in the visual or action-based context, such as consistent object movement and transitions in the main scene, related surroundings, and background. key changes in the visual or action-based context, such as object movement, color changes, scene transitions, or actions.
- **Action and Object Verification:** Reviewers should check that the objects and their corresponding actions mentioned in the modification text are present in both the input and target videos and that the text accurately describes their differences.

- **Temporal Alignment:** Ensure that the modification text aligns with the actual sequence of actions or events in the videos. The text should reflect changes in timing or flow between the two videos (e.g., one action follows another).
- **Comprehensiveness of Description:** The modification text should cover all relevant changes between the input and target video, providing a clear and complete description of the differences without leaving out important details.
- **Clarity and Conciseness:** Ensure that the modification text is clear, concise, and free of ambiguity. It should effectively communicate the modification in a straightforward and readable manner.
- **Cosine Similarity Threshold:** Use a cosine similarity threshold (e.g., 0.4) to identify low-quality modification texts that might require further human review and adjustment.
- **Statistical Tracking:** Maintain records of how many triplets were verified, how many corrections required, and what types of errors were most common. This helps refine the verification process over time. the number of verified triplets, required corrections, and types of common errors to refine the verification process over time.
- **Manual Corrections:** Reviewers can make manual corrections to the modification text if there are errors, missing details, or irrelevant information. They should focus on improving the alignment between text and video content.

Impact of Training Set Errors: Given the scale of the training set, it is reasonable to expect some inherent noise in the unverified portion. However, our multi-stage quality control ensures that errors are minimized and do not significantly impact the model’s performance. Empirically, we observe that training on denser, high-quality modification texts leads to significant improvements in retrieval accuracy (+3.4% Recall@1, as shown in Table 2 of the main paper). Furthermore, experiments on the manually verified subset demonstrate that even partial high-quality annotations substantially enhance the model’s generalization ability.

In summary, our dataset combines large-scale automated generation with meticulous human verification, ensuring that the test set is fully validated and the training set maintains a high standard of correctness. This makes Dense-WebVid-CoVR a robust and reliable benchmark for fine-grained composed video retrieval.

F. Dataset Statistics

As extra statistical analysis, in Fig. 4 we compare the distribution of the number of words in video descriptions between the WebVid-CoVR [5] dataset and our proposed Dense-WebVid-CoVR dataset. In the WebVid-CoVR dataset, the majority of video descriptions are short, typically containing between 3 and 10 words, with a sharp

As an intelligent video comprehension model, focus on these guidelines:

1. Differentiate recurring objects, count accurately, and identify movements and poses.
2. Understand directional movements and temporal order.
3. Pay attention to fine-grained actions with precision.
4. Assess incomplete actions without assuming completion.
5. Detect emotional, social, and visual cues.
6. Capture and analyze all relevant actions.
7. Identify unusual actions accurately.

Now, proceed with answering the following question in the form of detailed paragraph faithfully while keeping above guidelines in mind:
Question: What is happening in the video.

Figure 1. This figure presents an example prompt used to generate detailed video descriptions with the Gemini-1.5-Pro model. The prompt is designed to guide the model towards producing comprehensive, fine-grained descriptions by following specific guidelines. These guidelines instruct the model to accurately identify recurring objects, count them, capture directional movements, maintain temporal order, and assess actions with high precision. Additionally, the model is directed to detect emotional, social, and visual cues, ensuring a thorough understanding of the video’s content. This structured prompt enables the model to deliver rich, context-aware descriptions that are essential for high-quality video retrieval tasks.

System : "You are an AI designed to generate a modification_text based on two given captions (caption1 and caption2) that describe similar objects with some variations.

You will be provided with two captions: caption1 and caption2. Your task is to identify the key differences between these captions and describe the modification needed to transform caption1 into caption2. Consider the following example:
caption1: {}
caption2: {}
modification_text: {}"

User : "Now, generate the modification text for the following captions:
caption1: {}
caption2: {}
modification_text:

Write modification text as an instruction for AI model to retrieve video2 corresponding to caption2 from video1 with caption1 and modification_text.
Avoid making it a direct standalone caption of the target. Instead, highlight key transformations (objects, actions, spatial, or temporal changes) while ensuring multimodal reasoning. The modification text should enforce reliance on both video and text, preventing pure text-to-video retrieval.
Return modification_text in single sentence.
Do not hallucinate."

Figure 2. This figure illustrates an example prompt used with the GPT-4o model to generate precise modification texts between two video descriptions. The prompt instructs the model to analyze two given captions (caption1 and caption2), identify the key differences, and formulate a concise modification text to transform caption1 into caption2. The modification text is crafted as an instruction, guiding the retrieval model to locate the target video based on the described changes. By following these structured guidelines and avoiding any hallucinations, the prompt ensures that the generated modification text accurately reflects the required transformations, facilitating more precise and context-aware video retrieval.

decline beyond that. This limited description length of-

ten lacks the depth required for fine-grained video retrieval tasks. In contrast, our Dense-WebVid-CoVR dataset features significantly longer and more detailed video descriptions, with the distribution centered around 80 to 100 words per description. The increased richness and granularity of these descriptions enable more precise alignment between video content and modification text, greatly enhancing retrieval performance. This detailed annotation provides a stronger foundation for training retrieval models that can handle complex, fine-grained modifications in video content.

Additionally, Figure 5 illustrates the distribution of word counts in modification texts between the WebVid-CoVR [5] dataset and our proposed dataset. We can observe that in the WebVid-CoVR dataset, the majority of modification texts are very brief, with a word count predominantly ranging between 3 and 7 words. This brevity can limit the model’s ability to capture the distinct modifications needed for fine-grained video retrieval. In contrast, our dataset provides significantly more detailed modification texts, with the distribution centered around 30 to 50 words. The increased length and richness of our modification texts allow for a more comprehensive representation of the desired changes between videos, facilitating improved retrieval accuracy. By offering such detailed descriptions, our dataset enables models to handle complex modifications more effectively, leading to superior performance in fine-grained composed video retrieval tasks.

G. Qualitative Analysis

Figure 7 illustrates the impact of fine-grained modification texts by comparing the input query video and target retrieval video. It highlights our fine-grained modification text’s ability to capture subtle textual differences and retrieve the desired videos accurately. Figure 8 presents a comparison of datasets used in our study, emphasizing the diversity and richness of video-caption pairs in our proposed dataset, which contributes to the effective fine-grained retrieval performance. Figure 9 provides qualitative comparisons, showcasing retrieval outputs for various queries between Thawakar *et.al* [4] and our proposed approach on Dense-WebVid-CoVR set. These examples demonstrate that our proposed dataset and approach consistently aligns with the specified modifications, outperforming baseline methods in understanding and applying detailed textual inputs. This supplementary material further validates the robustness and reliability of our proposed method in fine-grained video retrieval.

Figure. 6 presents a qualitative between [4] and our approach on examples from CIRR validation set in the zero-shot setting. Compared to [4], our approach achieves favorable retrieval performance. For instance, the modification text in the second row is: *The target photo has three*

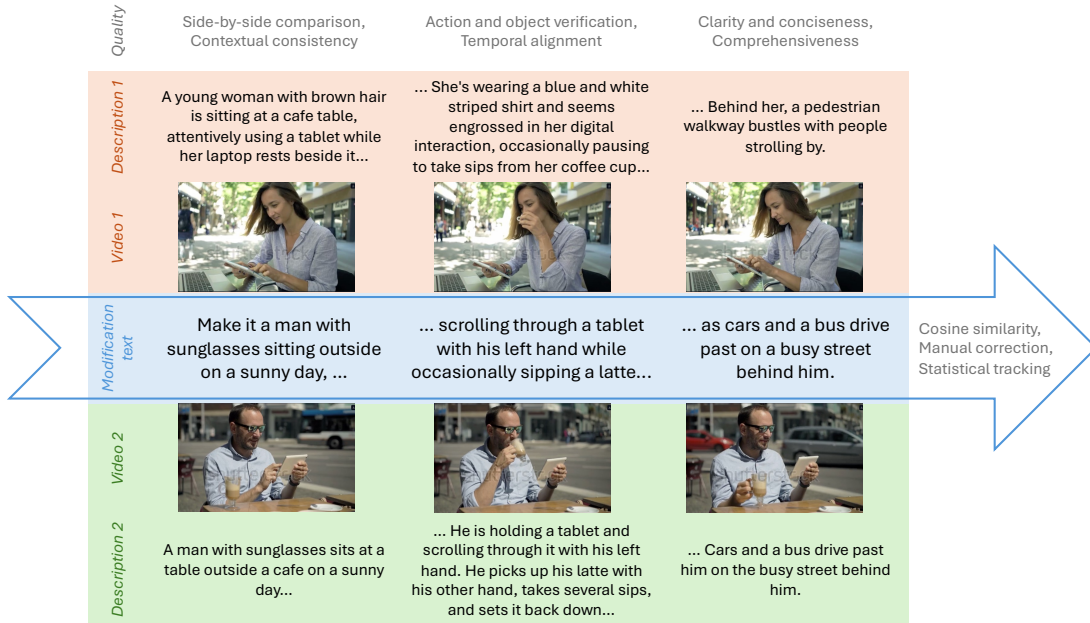


Figure 3. A data verification pipeline used for quality assessment of the generated data in the evaluation set. By ensuring a specific set of desired qualitative properties (in grey) of the textual data, we achieve high-quality video descriptions (in orange and green) and modification texts (in blue).

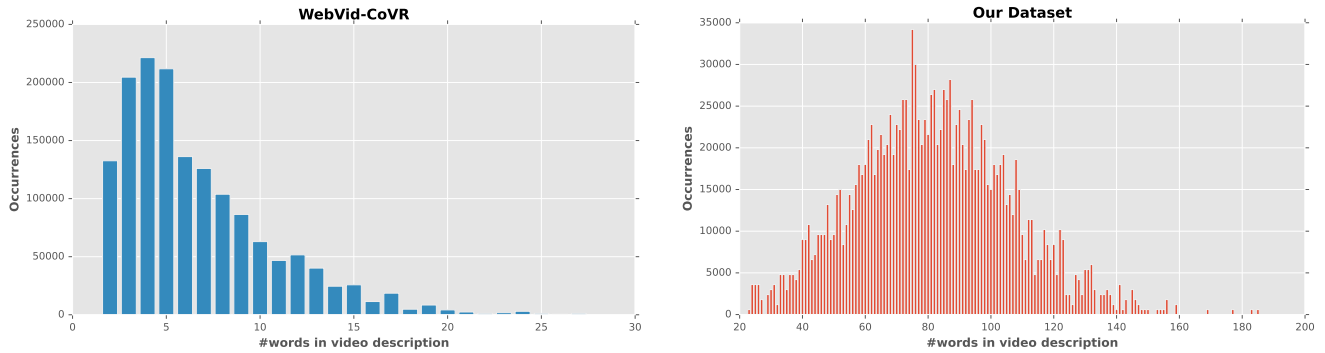


Figure 4. The comparison of the distribution of the number of words in video descriptions between the WebVid-CoVR [5] dataset (left) and our proposed dataset (right). In the WebVid-CoVR dataset, the majority of video descriptions are short, typically containing between 3 and 10 words, with a sharp decline beyond that. This limited description length often lacks the depth required for fine-grained video retrieval tasks. In contrast, our dataset features significantly longer and more detailed video descriptions, with the distribution centered around 80 to 100 words per description. The increased richness and granularity of these descriptions enable more precise alignment between video content and modification text, greatly enhancing retrieval performance. This detailed annotation provides a stronger foundation for training retrieval models that can handle complex, fine-grained modifications in video content.

antelopes in a wooded area looking at camera. Our approach provides accurate retrieval with respect to object count (three antelopes) and background context (wooded area). Our approach achieves accurate retrieval results with respect to the object count (three antelopes) and the back-

ground concept (wooded area).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko

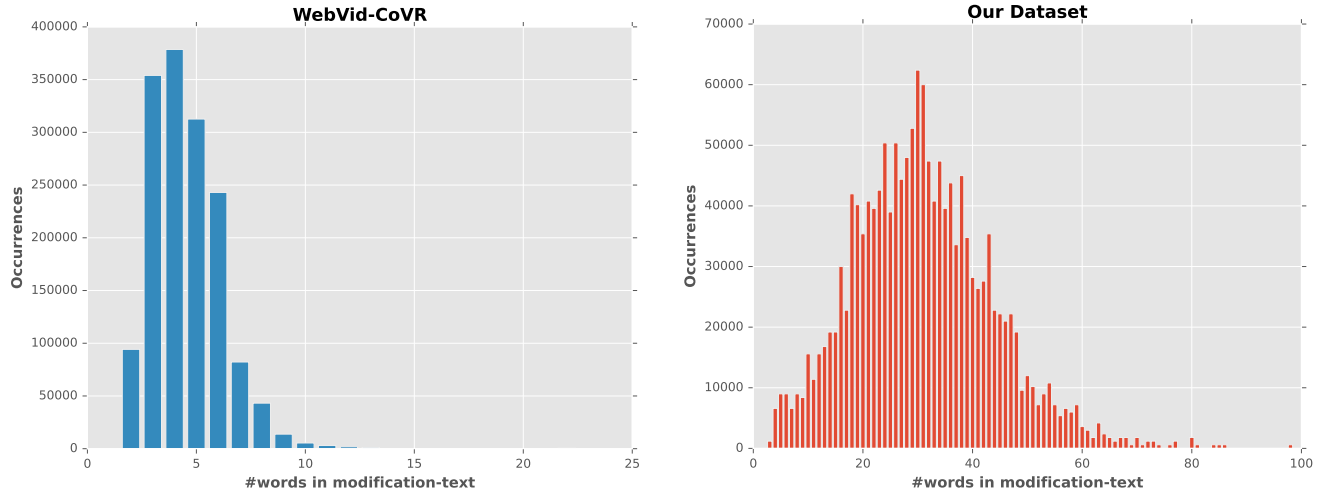


Figure 5. The graphs illustrate the distribution of word counts in modification texts between the WebVid-CoVR [5] dataset (left) and our proposed dataset (right). In the WebVid-CoVR dataset, the majority of modification texts are very brief, with a word count predominantly ranging between 3 and 7 words. This brevity can limit the model’s ability to capture the distinct modifications needed for fine-grained video retrieval. In contrast, our dataset provides significantly more detailed modification texts, with the distribution centered around 30 to 50 words. The increased length and richness of our modification texts allow for a more comprehensive representation of the desired changes between videos, facilitating improved retrieval accuracy. By offering such detailed descriptions, our dataset enables models to handle complex modifications more effectively, leading to superior performance in fine-grained composed video retrieval tasks.



Figure 6. Qualitative comparison between the recent CoVR model [4]) and our approach for zero-shot composed image retrieval (CoIR) task on CIRR validation set. Our unified fusion method captures the modification text more effectively, resulting in more accurate retrievals that align closely with the intended changes in the input query image. Best viewed zoomed in. Additional results are presented in suppl. material.

- [4] Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26896–26906, 2024. 1, 3, 5, 8
- [5] 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. 1, 2, 3, 4, 5

Alteneschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

- [2] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. *European Conference on Computer Vision (ECCV)*, 2024. 1
- [3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1



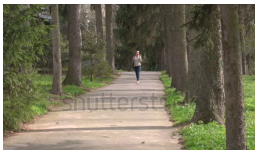







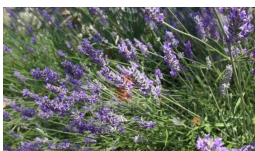









Input Video	Modification Text WebVid-CoVR	Modification Text Ours	Target Video
	Turn it into a lake	Transform the road in a snow-covered forest into a wide landscape with a calm river, bare trees, and sparse snowfall, ensuring the reflection in the water and serene stillness.	
	Turn into a jogger.	Transform into a scene with a gravel path, replace the woman with one in a teal tank top and black pants jogging towards the camera, and include other joggers along with green trees and partially obscured palm trees in the foreground.	
	make them pink.	change the cluster of white plumeria flowers to two pink and white plumeria flowers swaying in the breeze with a blurred greenish grey background.	
	replace with college	replace "VACATION" tag with "COLLEGE" one	
	Have her lounge	Have her relax on a beach chair at sunset, while facing away from the camera.	
	in the sunset	A breathtaking sunset over a picturesque lavender field with the setting sun casting a warm, golden glow, a distant field of sunflowers, a row of beehives near the sunflowers suggesting lavender honey production, and captivating atmosphere.	
	have him pour a cup of coffee	Change the focus from a man enjoying espresso in nature to a coffee maker brewing and pouring coffee into a red mug on a white countertop.	
	Make them run from the sun	Transform the scene to a sunset with serene, gently drifting clouds and a glowing sun, utilizing a warm color palette and evoking a tranquil atmosphere.	
	change to female	Change to a woman with long dark hair, wearing a green shirt, already wearing a light blue surgical mask, in a forest with a visible path, static video with no movement.	
	make it a harvesting scene	edit : Transform the scene from a calm, wind-swept wheat field to an active wheat harvest with a yellow combine harvester working under a blue sky with scattered white clouds.	

Figure 7. The modification-text comparison between WebVid-CoVR and our improved Dense-WebVid-CoVR dataset.



















Input (Video-1)	Target (Video-2)	WebVid Description (Video-1)	WebVid Description (Video-2)	Modification Text (WebVid)	Our Description (Video-1)	Our Description (Video-2)	ModificationText (Ours)
		Clouds and blue sky.	Grassland and blue sky	put a grassland background	The video shows a timelapse of fluffy white cumulus clouds moving across a blue sky. The clouds appear to be moving from left to right, expanding and changing shape as they drift. The sunlight illuminates the clouds, creating a sense of depth and movement.	The video displays a serene outdoor scene with a lone tree standing on a grassy hill against a backdrop of a bright blue sky dotted with fluffy white clouds. The tree's leaves rustle gently in the wind, creating a sense of tranquility and peace. The grass on the hill is lush and green,	Add a serene outdoor scene with a lone tree on a grassy hill, and make the camera static to capture the subtle movements of the tree and clouds, evoking tranquility and the beauty of nature.
		Squeezing a fresh and juicy orange. extreme close up.	Squeezing a juicy tomato. extreme close up.	make a tomato	The video provides a close-up view of the juicy flesh of an orange. The segments of the orange are visible, filled with juice that reflects light. The viewpoint	The video presents an extreme close-up, macro view of a sliced tomato. The vibrant red flesh of the tomato is visible, glistening with moisture, highlighting its juicy nature. Numerous yellow seeds	make it a sliced tomato with vibrant red flesh, yellow seeds, and a subtle flow of juice creating a dynamic element within an extreme close-up, macro view.
		Birch leaves	Water leaves	change the leaves to water	A gentle breeze rustles the leaves of a birch tree, creating a soothing, natural ambiance. The slender branches, adorned with vibrant green and delicate brown leaves, sway gracefully in the wind. The background, a soft blur of green grass and subtle hints of yellow flowers, adds to the tranquil atmosphere. The camera remains static, focusing on the intricate dance	The video shows a dark, rippling water surface, likely filmed from above. Three yellow leaves float on the surface, drifting slightly with the movement of the water. The leaves appear to be spread out from each other, not clumped together. The water is in constant motion, with small waves continuously rippling	Change the scene from a birch tree with rustling leaves in a tranquil natural setting to a dark, rippling water surface with three scattered yellow leaves drifting on it.
		Flight over the mountains	Flight over the mist	Cover the mountains with mist	The video shows a picturesque aerial view of rolling hills covered in golden-yellow grass. The camera moves slowly over the landscape, providing a sweeping panorama of the undulating terrain. Patches of dark green trees are scattered amongst the hills, adding a touch of contrast to the scene. Beyond the foreground hills, a range of larger mountains rises majestically	The video captures a breathtaking aerial view of a lush green forest partially enveloped in a thick blanket of fog. The drone camera points directly downwards, providing a top-down perspective of the verdant landscape. Initially, the scene is dominated by swirling fog, obscuring the forest floor. However, as the video progresses, the fog gradually dissipates, revealing more of the underlying	Change the rolling hills covered in golden-yellow grass, patches of dark green trees, and larger mountains under a clear blue sky to a lush green forest, partially enveloped in thick fog with a winding dirt road flanked by dense clusters of trees emerging as the fog dissipates.
		Young man relaxing with coffee on gazebo bed in garden	Young man relaxing with drink on gazebo bed in garden	Have a drink instead	A man is comfortably reclining on a daybed adorned with vibrant blue and yellow cushions, set within a charming bamboo gazebo surrounded by lush greenery. He is initially engrossed in his laptop, his bare feet relaxed. After a while, he closes the laptop, stretches leisurely, and reaches for a cup of coffee on	A man is relaxing on a daybed in a gazebo, casually dressed in a white shirt and gray shorts. He appears to be enjoying a refreshing drink, taking sips from a glass while observing his serene surroundings. The gazebo is nestled amidst lush greenery, ...	Change man's attire, replace laptop and coffee with a glass of refreshing drink, and simplify the description of the setting.
		A tree with alpha perfect for compositing	Palm tree in the breeze with alpha perfect for compositing	Replace with palm tree	The video displays a static image of a lush green tree with red berries against a plain black background. The tree remains completely motionless throughout the duration of the footage, showing no signs of movement or alteration in its pose. ...	The video showcases a solitary palm tree set against a stark black background. The palm remains stationary throughout the video, exhibiting no movement or change in its pose. The leaves of the palm exhibit a gentle swaying motion, indicative of a light breeze. The video focuses solely on the visual aspects of the palm,	Replace lush green tree with red berries and gray rock with a solitary palm tree, and add gentle leaf movement to the palm leaves.
		Slow motion falling apples	Slow motion falling peppers	replace with peppers	The video shows three apples, two red and one green, falling onto a wet black surface. The apples drop from out of the frame, descending from the top of the scene. The green apple is in the center, slightly ahead of the two red apples on either side. Upon impacting the surface, all three apples create splashes of water, indicating the surface was wet beforehand.	The video showcases a trio of bell peppers red, orange, and yellow dynamically interacting with water splashes on a sleek black surface. The peppers fall from above, landing on the reflective surface. As they settle, a surge of water erupts around them, momentarily obscuring their forms. The peppers, slick with water droplets, slightly shift their positions due to	replace apples with bell peppers of red, orange, and yellow.
		Waves at sunrise on the beach anse lazio. island of prasin in seychelles.	Boat at sunrise on the beach anse lazio. island of prasin in seychelles.	turn it into a boat	The video showcases a pristine beach with white sand and turquoise water. Waves, adorned with whitecaps, continuously roll onto the shore, creating a dynamic and mesmerizing scene. The camera remains stationary, capturing the rhythmic ebb and flow of the waves as they crash and recede. A lush, green, tree-covered hill flanks one side of the beach, providing a ...	A white catamaran rests peacefully on the tranquil turquoise waters, its white sail furled, likely anchored just offshore. Gentle waves, crested with whitecaps, roll rhythmically toward the shore, creating a soothing soundscape. The vibrant blue sky is adorned with an array of puffy white cumulus clouds, suggesting a bright and sunny day.	Replace the pristine beach and tree-covered hill with a white catamaran and idyllic islands.
		Red coffee cup on wooden table. dolly shot	Yellow coffee cup on wooden table. dolly shot	make it yellow	The video showcases a still image of a red coffee cup filled with a creamy beverage and adorned with chocolate sprinkles. The cup sits atop a matching red saucer, resting on a wooden tray positioned on a dark table. A silver spoon accompanies the cup, placed to its left. The backdrop consists of a vibrant scene of lush ...	A yellow cup filled with creamy coffee sits on a yellow saucer upon a wooden tray atop a brown wooden table. A silver teaspoon lies to the left of the cup on the tray. The backdrop is a lush green tropical garden, slightly out of focus. The coffee remains ...	Make the coffee cup and saucer yellow, change the table color to brown, and describe the foliage as a lush green tropical garden.

Figure 8. Data samples comparison from WebVid-CoVR our improved Dense-WebVid-CoVR dataset.

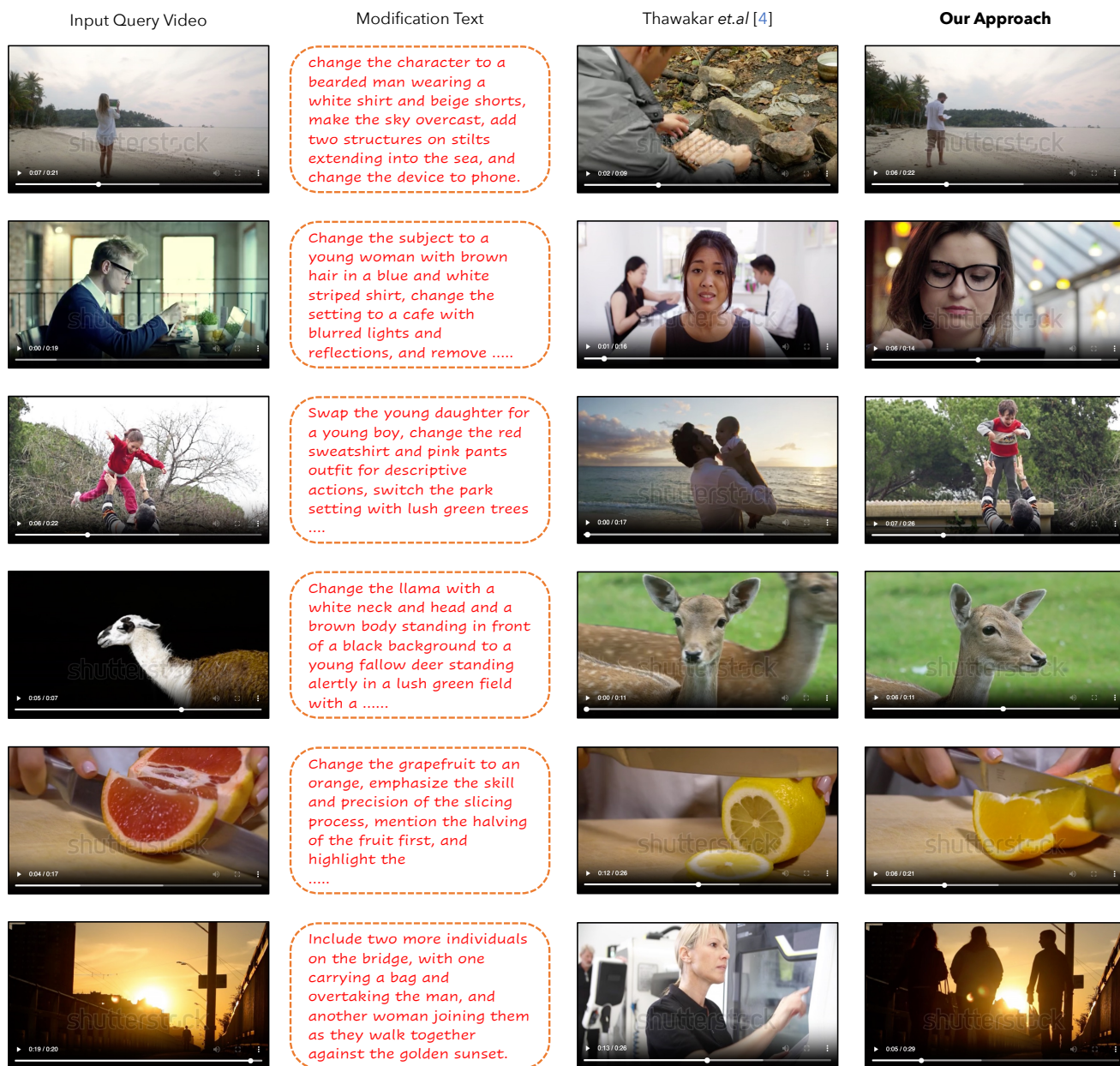


Figure 9. The figure demonstrates the effectiveness of fine-grained modification text in video retrieval, comparing the proposed approach with Thawakar *et al.* [4]. Each row showcases a query video, a detailed modification text specifying changes in subjects, actions, objects, or scenes, and the results retrieved by both methods. The proposed approach consistently retrieves videos that accurately reflect the described modifications, such as changes in characters, settings, objects, or activities, while the comparison method often fails to adapt to the nuanced details. This highlights the proposed method’s superior ability to interpret and apply fine-grained textual instructions for precise video retrieval.