Composed Video Retrieval via Enriched Context and Discriminative Embeddings
(Supplementary)

A. Description Generation

We used recent open-source multi-model conversational model such as MiniGPT-4 [3] to generate the detailed description of input video using middle frame of the video. We utilize pre-determined prompts in the given format $\#\#\#Human :< Img >< ImageFeature >< /Img >< Instruction > \#\#\#Assistant :$ ###Human, corresponds to the prompt. ###Assistant, serves the purpose of determining the system role, which in our case is defined as ”*You will be able to see the image once I provide it to you. Please answer my questions.*.” $< Instruction >$ refers to a the instruction given to LLM which is ”Describe the given image in details”. Both text queries undergo tokenization and fed into the LLM [1] which generates the detailed description of the input image.

Fig. 1 and 2 plays a crucial role in emphasizing the significance of the detailed captions generated by our method for effective target video retrieval. The figure distinctly illustrates the limitations of the default short captions used in WebVid dataset. These conventional captions often lack the essential contextual details needed to accurately matched with the corresponding target videos. By contrast, our generated detailed captions encompass a broader range of contextual elements, ensuring that the key aspects of the reference video are effectively preserved and mirrored in the retrieved target videos. This comparison in both the figures 1 and 2 not only highlights the deficiencies in webvid captions but also underscores the enhanced performance and accuracy achieved through our approach using these detailed descriptions. The examples serves as a compelling visual representation of the added value brought by our detailed descriptions for composed video retrieval.

Problem with generating descriptions from multi-modal conversational LLM is it’s hallucinated outputs. These hallucinations refer to the occurrences where the model generates responses or outputs that are either irrelevant or factually incorrect, deviating from the expected or logical outcome based on the input data. The process of identifying these hallucinatory examples is detailed in Section 3.2 of the paper. In this section, we describe the methodology used to detect and categorize these instances. Our approach involves setting a threshold for hallucination detection, which, in this case, is determined to be 0.2. This threshold represents a quantifiable measure or criterion used to filter out responses that are likely to be hallucinations.

Once these examples are filtered based on the hallucination threshold, we then undertake a manual review and correction process. This manual intervention is crucial as it not only helps in rectifying the inaccuracies in the model’s outputs but also provides valuable insights into the nature and characteristics of the hallucinations produced by the model. Through this correction process, we can refine the model’s performance, enhancing its reliability and accuracy.

B. Additional Qualitative Results

Fig. 3 and 4 shows the additional qualitative comparison with our baseline CoVR-BLIP [2] where the top row represents the baseline [2] results and bottom row represents the our approach. For each example video, top-10 retrieved videos with their similarity scores were shown. Our approach correctly retrieved targeted video as highlighted in green tic whereas the baseline CoVR-BLIP [2] not able to retrieve correct target video with the highest similarity score.

Figures 3 and 4 illustrate a qualitative comparison between baseline CoVR-BLIP [2] and our approach. These figures highlights the advancements and improvements our method offers over the baseline in the context of composed video retrieval accuracy and relevance. In both figures, the top row displays results obtained using the baseline CoVR-BLIP model, while the bottom row showcases the outcomes achieved with our approach. This side-by-side arrangement allows for a direct and intuitive comparison between the two methods. For each example video featured in these figures, the top-10 retrieved videos, along with their respective similarity scores, are presented. These similarity scores are key indicators of how well the retrieved videos match the input query video in terms of content, context, and change elements.

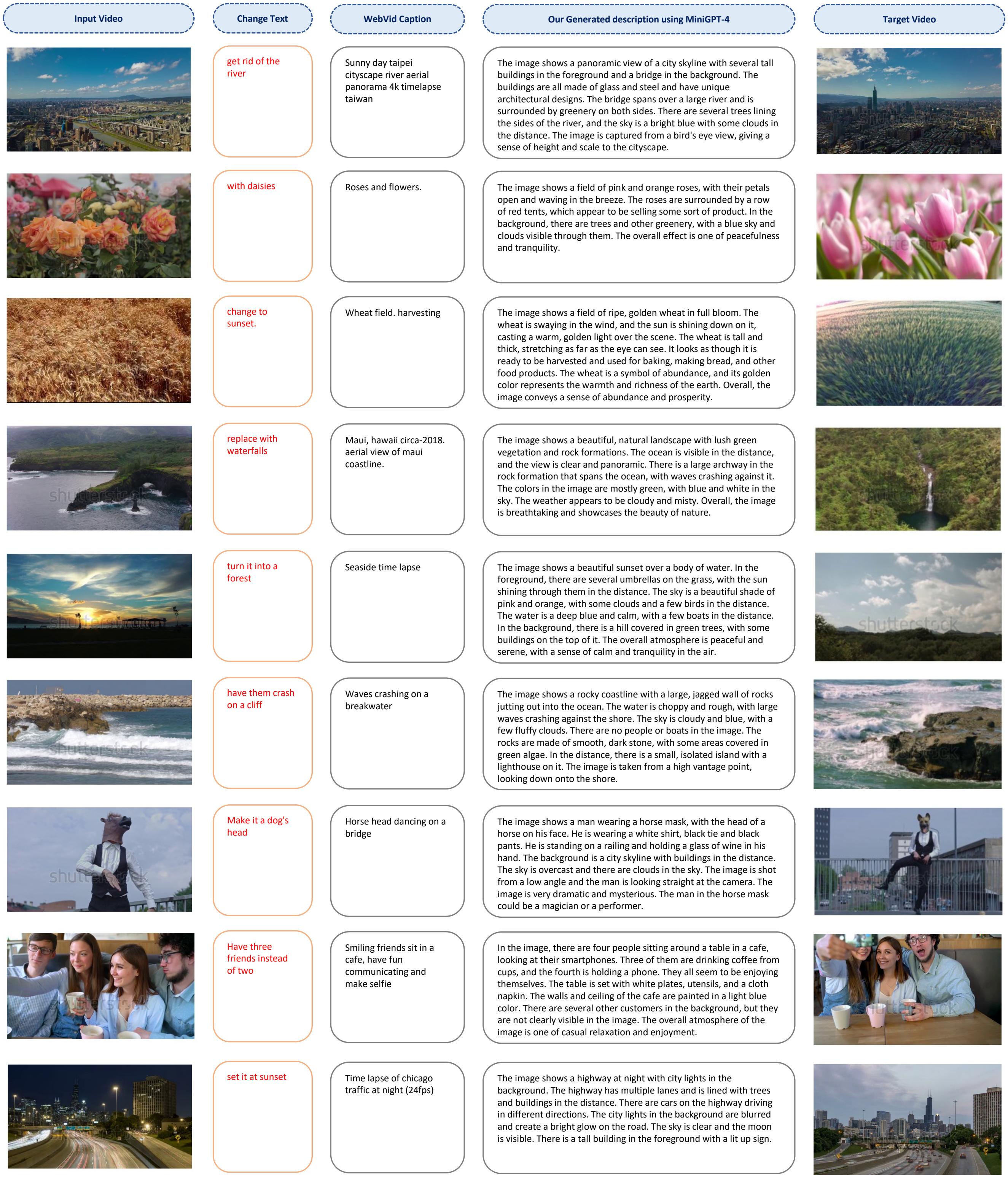
A significant aspect highlighted in these figures is the efficacy of our approach in accurately retrieving the targeted video. This is visually represented by a green tick mark, indicating instances where our model successfully identified the correct target video with the highest similarity score. In contrast, the baseline CoVR-BLIP model often fails to retrieve the correct target video with the highest similarity score. This difference is critical as it underscores the enhanced precision and effectiveness of our method in identifying and ranking relevant video content in response to a given input. With these comparative examples, Figures 3 and 4 not only demonstrate the superiority of our approach over the baseline but also offer tangible evidence of the practical improvements our model brings to the field of composed video retrieval using detailed descriptions. This comparison is useful in showcasing the real-world applicability and benefits of the advancements our approach contributes to the domain. Additionally, Fig. 5 shows the qualitative examples for our proposed approach where our approach successfully retrieve the correct target video as highest similarity.

References

[1] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[2] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *AAAI*, 2024. 1, 5, 6

[3] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1



Input Video	Change Text	WebVid Caption	Our Generated description using MiniGPT-4	Target Video
	get rid of the river	Sunny day taipei cityscape river aerial panorama 4k timelapse taiwan	The image shows a panoramic view of a city skyline with several tall buildings in the foreground and a bridge in the background. The buildings are all made of glass and steel and have unique architectural designs. The bridge spans over a large river and is surrounded by greenery on both sides. There are several trees lining the sides of the river, and the sky is a bright blue with some clouds in the distance. The image is captured from a bird's eye view, giving a sense of height and scale to the cityscape.	
	with daisies	Roses and flowers.	The image shows a field of pink and orange roses, with their petals open and waving in the breeze. The roses are surrounded by a row of red tents, which appear to be selling some sort of product. In the background, there are trees and other greenery, with a blue sky and clouds visible through them. The overall effect is one of peacefulness and tranquility.	
	change to sunset.	Wheat field. harvesting	The image shows a field of ripe, golden wheat in full bloom. The wheat is swaying in the wind, and the sun is shining down on it, casting a warm, golden light over the scene. The wheat is tall and thick, stretching as far as the eye can see. It looks as though it is ready to be harvested and used for baking, making bread, and other food products. The wheat is a symbol of abundance, and its golden color represents the warmth and richness of the earth. Overall, the image conveys a sense of abundance and prosperity.	
	replace with waterfalls	Maui, hawaii circa-2018. aerial view of maui coastline.	The image shows a beautiful, natural landscape with lush green vegetation and rock formations. The ocean is visible in the distance, and the view is clear and panoramic. There is a large archway in the rock formation that spans the ocean, with waves crashing against it. The colors in the image are mostly green, with blue and white in the sky. The weather appears to be cloudy and misty. Overall, the image is breathtaking and showcases the beauty of nature.	
	turn it into a forest	Seaside time lapse	The image shows a beautiful sunset over a body of water. In the foreground, there are several umbrellas on the grass, with the sun shining through them in the distance. The sky is a beautiful shade of pink and orange, with some clouds and a few birds in the distance. The water is a deep blue and calm, with a few boats in the distance. In the background, there is a hill covered in green trees, with some buildings on the top of it. The overall atmosphere is peaceful and serene, with a sense of calm and tranquility in the air.	
	have them crash on a cliff	Waves crashing on a breakwater	The image shows a rocky coastline with a large, jagged wall of rocks jutting out into the ocean. The water is choppy and rough, with large waves crashing against the shore. The sky is cloudy and blue, with a few fluffy clouds. There are no people or boats in the image. The rocks are made of smooth, dark stone, with some areas covered in green algae. In the distance, there is a small, isolated island with a lighthouse on it. The image is taken from a high vantage point, looking down onto the shore.	
	Make it a dog's head	Horse head dancing on a bridge	The image shows a man wearing a horse mask, with the head of a horse on his face. He is wearing a white shirt, black tie and black pants. He is standing on a railing and holding a glass of wine in his hand. The background is a city skyline with buildings in the distance. The sky is overcast and there are clouds in the sky. The image is shot from a low angle and the man is looking straight at the camera. The image is very dramatic and mysterious. The man in the horse mask could be a magician or a performer.	
	Have three friends instead of two	Smiling friends sit in a cafe, have fun communicating and make selfie	In the image, there are four people sitting around a table in a cafe, looking at their smartphones. Three of them are drinking coffee from cups, and the fourth is holding a phone. They all seem to be enjoying themselves. The table is set with white plates, utensils, and a cloth napkin. The walls and ceiling of the cafe are painted in a light blue color. There are several other customers in the background, but they are not clearly visible in the image. The overall atmosphere of the image is one of casual relaxation and enjoyment.	
	set it at sunset	Time lapse of chicago traffic at night (24fps)	The image shows a highway at night with city lights in the background. The highway has multiple lanes and is lined with trees and buildings in the distance. There are cars on the highway driving in different directions. The city lights in the background are blurred and create a bright glow on the road. The sky is clear and the moon is visible. There is a tall building in the foreground with a lit up sign.	

Figure 1. Examples of our detailed descriptions providing richer information for reference video compared to default Webvid-CoVR Captions. Our detailed descriptions enhance the understanding and interpretation of the reference video, offering more context, clarity, and specific information that the default webvid captions may lack.

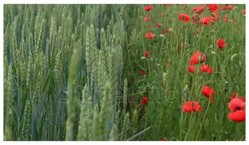
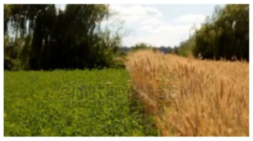
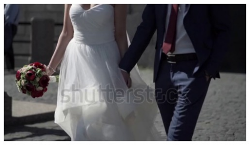
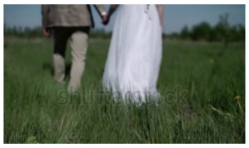
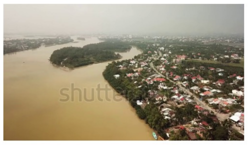
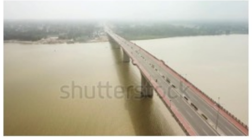
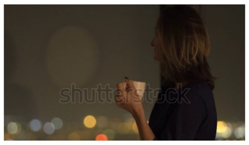
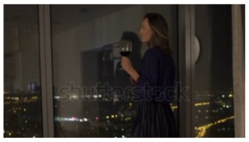
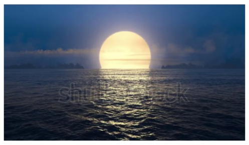
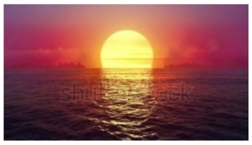
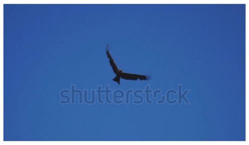
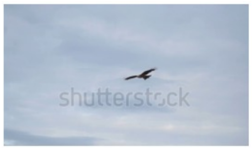
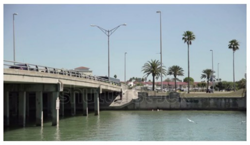
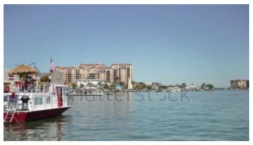
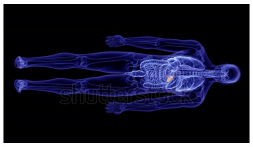
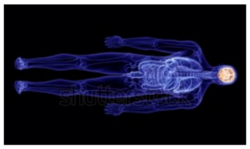
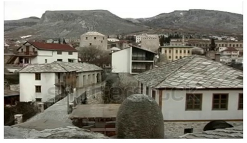
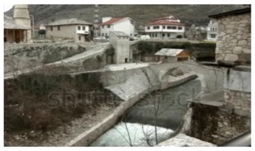
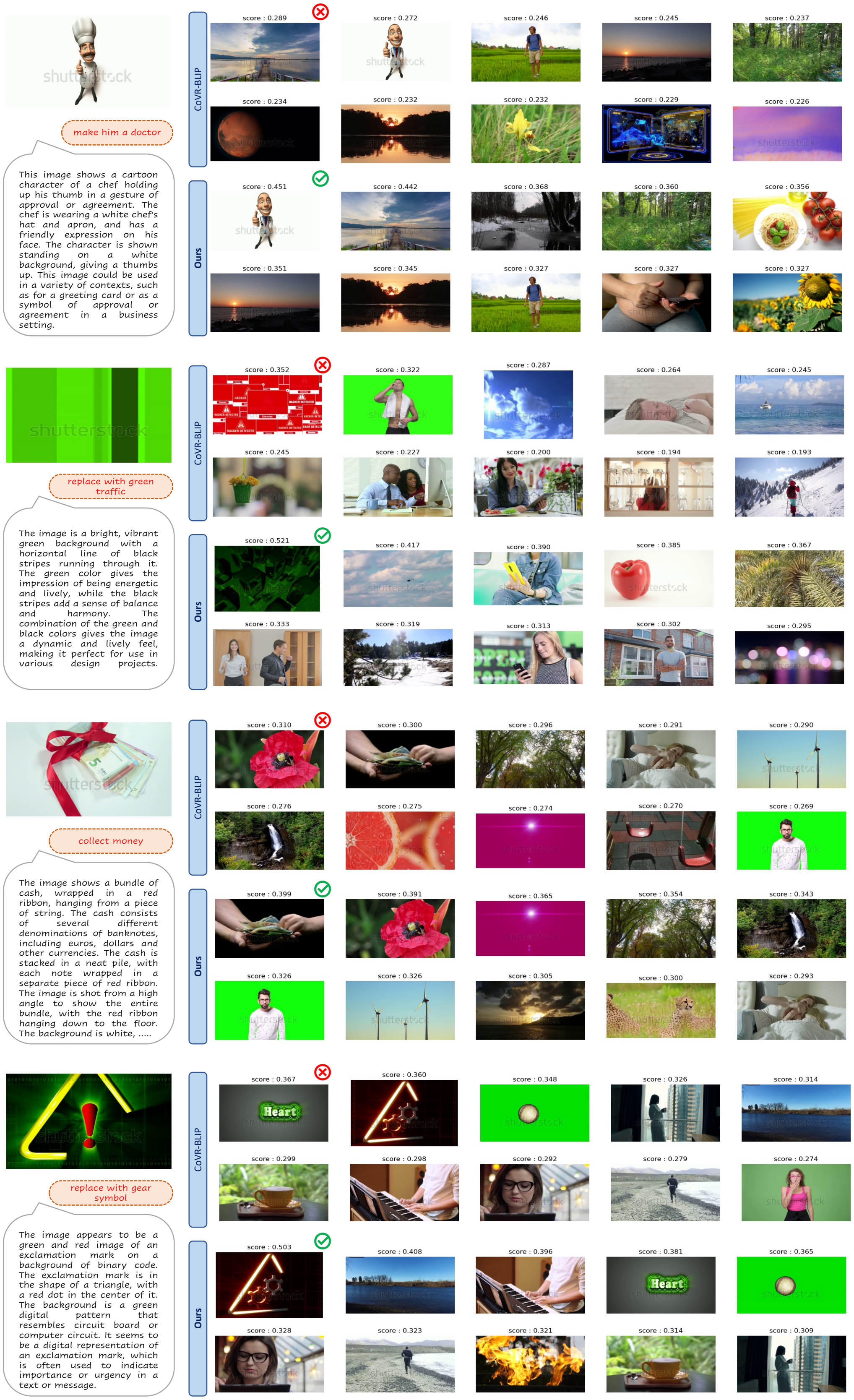
Input Video	Change Text	WebVid Caption	Our Generated description using MiniGPT-4	Target Video
	Wind on a wheat field	Poppy on a wheat field	The image shows a field of tall, green grass and red poppies growing in it. The grass is waving in the breeze and the poppies are in full bloom, their bright red petals swaying in the wind. The sky is clear and blue, with a few white clouds floating overhead. The sun is shining down on the scene, casting dappled shadows on the grass. In the distance, there is a small farmhouse with a red roof and white picket fence. The farmhouse looks like it is surrounded by trees and has a small barn in the back.	
	In the field	Lovely couple walking in a city slowmotion	The image shows a bride and groom walking down a street on their wedding day. The bride is wearing a white dress and holding a bouquet of red roses. The groom is wearing a black tuxedo and holding her hand. They are walking down the street holding hands and looking at each other with smiles on their faces. Behind them, there is a stone wall with trees and shrubs growing on it. The scene is bathed in warm sunlight, casting shadows on the wall. In the background, there are buildings and cars driving on the street.	
	change to be a bridge	Aerial view of village in hoi an,vietnam	1. The image shows a small town nestled along the banks of a river with trees and houses lining the sides of the river.2. The town appears to be quite old and quaint with wooden houses and a few modern buildings mixed in. The town is surrounded by green trees and bushes, giving it a serene and peaceful atmosphere.3. There are a few boats tied up to the banks of the river, suggesting that the town has a riverfront. The river is fairly wide and appears to have some water flowing through it.	
	Drink wine instead of tea	Young woman drinking tea and admire view from window at home at night	The image shows a woman standing at a window in a modern office building, looking out at the city at night. She is holding a mug of hot coffee and smiling at the view. The background is lit up by city lights, and there are some trees and buildings visible in the distance. The woman is wearing a dark sweater and jeans, and her hair is dark and casual. The overall feeling of the image is one of warmth and contentment.	
	make it more fantasy-like	Seamlessly looped flight over ocean at blue sunset.	The image depicts a sunset over a vast ocean with a cloudy sky in the background. The sun is setting behind the clouds, casting a warm glow on the water. The colors in the sky are shades of orange and pink, with a slight purple tint to them. The water is calm, with a few small waves lapping at the shore. In the foreground, there are some trees on the shore, their branches swaying in the breeze. The overall mood of the image is peaceful and calming.	
	make it cloudy	Bird of prey flying against the blue sky	This image shows a large bird flying in the blue sky. The bird has a long black tail and a large wingspan. Its body is brown with black feathers and its wings are spread open as it flies. The background is a bright blue sky with a few clouds in the distance. The image is quite detailed and allows for a clear view of the bird's anatomy and flight movements.	
	Make it a dog's head	Clearwater, united states - may, 2017: bridge in clearwater	The image shows a large bridge spanning a body of water with several cars driving under it. On either side of the bridge are tall buildings with palm trees in the foreground. The sky is clear and blue with a few clouds in the distance. The water is calm and reflective in the sunlight. There are no people or boats visible in the image. The bridge appears to be made of concrete and steel.Overall, the image is a view of a large, modern bridge spanning a body of water with tall buildings in the background.	
	Change to the larynx	3d rendered, medically accurate illustration of the human gallbladder	The image shows a human skeleton with an X-ray view of the body's organs and bones. The skeleton appears to be a male adult, with a head and limbs visible. The image is in a blue color, with no other visible features or background. The organs and bones appear clear and visible, with the heart and lungs in the chest area, the liver and spleen in the upper abdomen, and the kidneys and adrenal glands in the lower abdomen.	
	Have a waterfall	Roofs mostar	This image shows a small town with old buildings and a rocky terrain in the background. The buildings are made of stone and have wooden shutters on the windows. The roofs are made of clay tiles and there are mountains in the background. There are no trees or other vegetation in the foreground, only rocks and the buildings. The town seems to be located in a mountainous area with steep slopes and rocky terrain. There is no road or other infrastructure visible in the image.	

Figure 2. Additional examples of our detailed descriptions providing richer information for reference video compared to default Webvid-CoVR Captions. Our detailed descriptions enhance the understanding and interpretation of the reference video, offering more context, clarity, and specific information that the default webvid captions may lack.



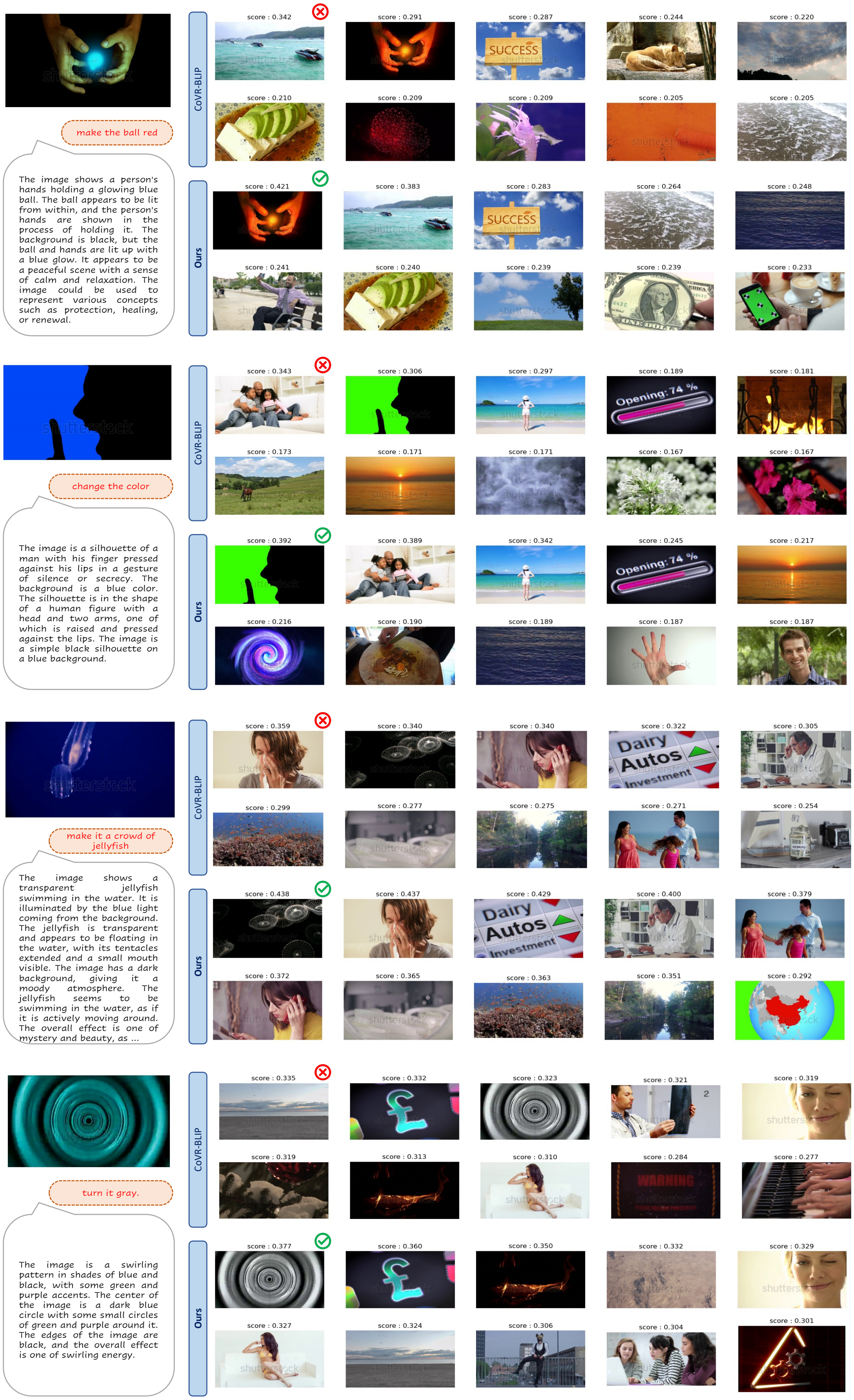
This image shows a cartoon character of a chef holding up his thumb in a gesture of approval or agreement. The chef is wearing a white chef's hat and apron, and has a friendly expression on his face. The character is shown standing on a white background, giving a thumbs up. This image could be used in a variety of contexts, such as for a greeting card or as a symbol of approval or agreement in a business setting.

The image is a bright, vibrant green background with a horizontal line of black stripes running through it. The green color gives the impression of being energetic and lively, while the black stripes add a sense of balance and harmony. The combination of the green and black colors gives the image a dynamic and lively feel, making it perfect for use in various design projects.

The image shows a bundle of cash, wrapped in a red ribbon, hanging from a piece of string. The cash consists of several different denominations of banknotes, including euros, dollars and other currencies. The cash is stacked in a neat pile, with each note wrapped in a separate piece of red ribbon. The image is shot from a high angle to show the entire bundle, with the red ribbon hanging down to the floor. The background is white, ...

The image appears to be a green and red image of an exclamation mark on a background of binary code. The exclamation mark is in the shape of a triangle, with a red dot in the center of it. The background is a green digital pattern that resembles circuit board or computer circuit. It seems to be a digital representation of an exclamation mark, which is often used to indicate importance or urgency in a text or message.

Figure 3. The comparison between the baseline CoVR-BLIP [2] and our approach on selected video samples from the WebVid-CoVR test set. The correct target video retrieved is highlighted with green marks with highest similarity score. The results demonstrates the superior accuracy and relevance of the retrieved videos using our method in comparison with baseline [2] for a given input video, highlighting the significant improvements and refinements our approach brings to video content analysis and retrieval.



The image shows a person's hands holding a glowing blue ball. The ball appears to be lit from within, and the person's hands are shown in the process of holding it. The background is black, but the ball and hands are lit up with a blue glow. It appears to be a peaceful scene with a sense of calm and relaxation. The image could be used to represent various concepts such as protection, healing, or renewal.

The image is a silhouette of a man with his finger pressed against his lips in a gesture of silence or secrecy. The background is a blue color. The silhouette is in the shape of a human figure with a head and two arms, one of which is raised and pressed against the lips. The image is a simple black silhouette on a blue background.

The image shows a transparent jellyfish swimming in the water. It is illuminated by the blue light coming from the background. The jellyfish is transparent and appears to be floating in the water, with its tentacles extended and a small mouth visible. The image has a dark background, giving it a moody atmosphere. The jellyfish seems to be swimming in the water, as if it is actively moving around. The overall effect is one of mystery and beauty, as ...

The image is a swirling pattern in shades of blue and black, with some green and purple accents. The center of the image is a dark blue circle with some small circles of green and purple around it. The edges of the image are black, and the overall effect is one of swirling energy.

Figure 4. Our approach, as illustrated in the comparison with the baseline CoVR-BLIP [2], showcases its efficacy on chosen video samples from the WebVid-CoVR test set. Videos retrieved by our method that precisely match the target are marked in green, indicating the highest similarity scores.

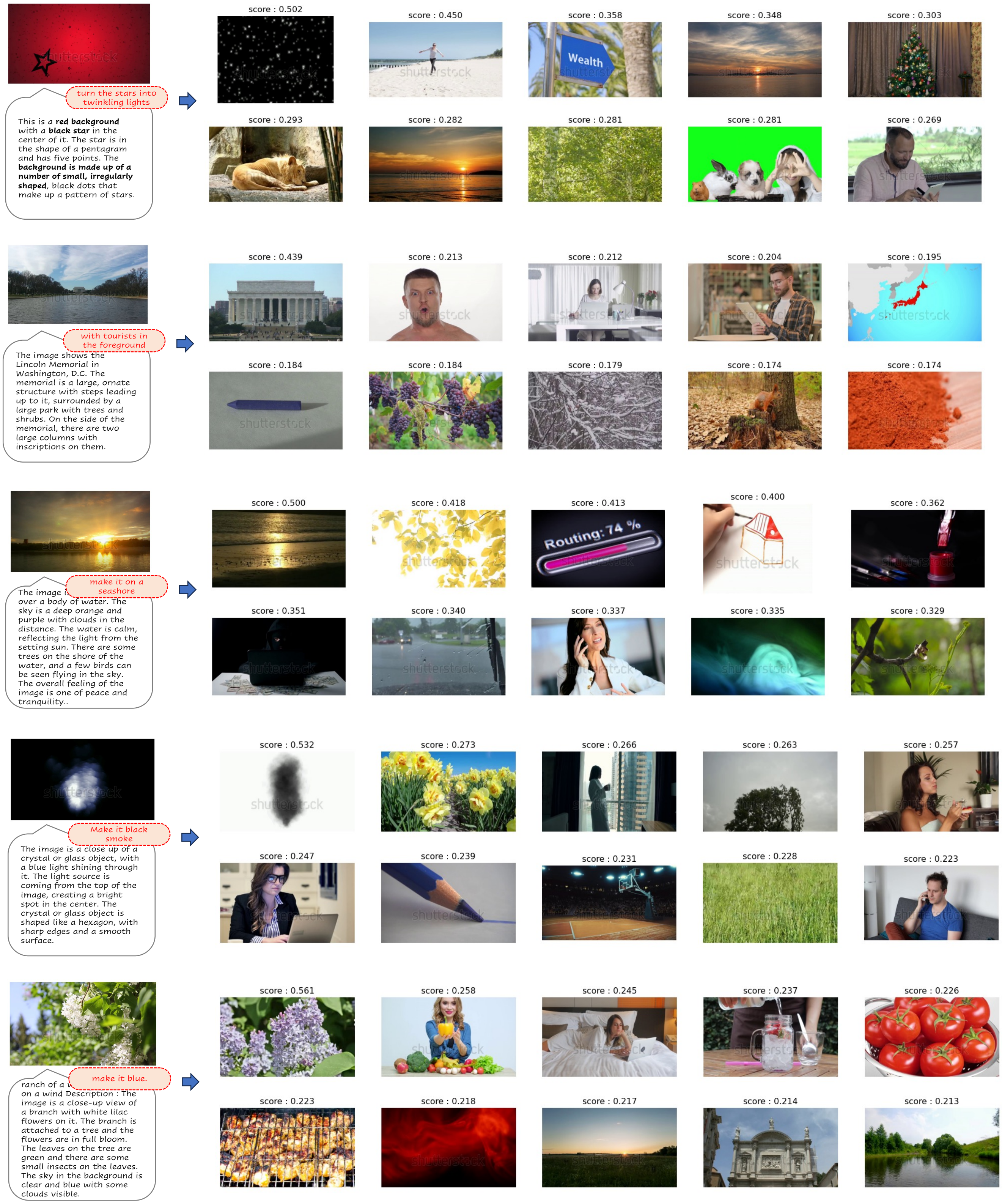

turn the stars into twinkling lights
This is a red background with a black star in the center of it. The star is in the shape of a pentagram and has five points. The background is made up of a number of small, irregularly shaped, black dots that make up a pattern of stars.
score : 0.502
score : 0.450
score : 0.358
Wealth
score : 0.348
score : 0.303
score : 0.293
score : 0.282
score : 0.281
score : 0.281
score : 0.269
with tourists in the foreground
The image shows the Lincoln Memorial in Washington, D.C. The memorial is a large, ornate structure with steps leading up to it, surrounded by a large park with trees and shrubs. On the side of the memorial, there are two large columns with inscriptions on them.
score : 0.439
score : 0.213
score : 0.212
score : 0.204
score : 0.195
score : 0.184
score : 0.184
score : 0.179
score : 0.174
score : 0.174
make it on a seashore
The image is over a body of water. The sky is a deep orange and purple with clouds in the distance. The water is calm, reflecting the light from the setting sun. There are some trees on the shore of the water, and a few birds can be seen flying in the sky. The overall feeling of the image is one of peace and tranquility..
score : 0.500
score : 0.418
score : 0.413
Routing: 74 %
score : 0.400
score : 0.362
score : 0.351
score : 0.340
score : 0.337
score : 0.335
score : 0.329
Make it black smoke
The image is a close up of a crystal or glass object, with a blue light shining through it. The light source is coming from the top of the image, creating a bright spot in the center. The crystal or glass object is shaped like a hexagon, with sharp edges and a smooth surface.
score : 0.532
score : 0.273
score : 0.266
score : 0.263
score : 0.257
score : 0.247
score : 0.239
score : 0.231
score : 0.228
score : 0.223
make it blue.
ranch of a w on a wind Description : The image is a close-up view of a branch with white lilac flowers on it. The branch is attached to a tree and the flowers are in full bloom. The leaves on the tree are green and there are some small insects on the leaves. The sky in the background is clear and blue with some clouds visible.
score : 0.561
score : 0.258
score : 0.245
score : 0.237
score : 0.226
score : 0.223
score : 0.218
score : 0.217
score : 0.214
score : 0.213

Figure 5. Positive Outcomes Demonstrated by our approach on WebVid-CoVR Test Set. The figure presents the top-10 videos retrieved for an input video, showcasing the effectiveness of our approach. Each video is accompanied by a similarity score, indicating the degree of correlation between the input composition and the retrieved target video, thus emphasizing the precision and relevance of our retrieval system in identifying closely matched content.