
Supplementary Material:

CoVR: Learning Composed Video Retrieval from Web Video Captions

Anonymous Author(s)

Affiliation

Address

email

1 This document provides dataset statistics (Section A), implementation details (Section B), additional
2 experiments (Section C), and qualitative examples (Section D). We also provide the code (code.zip),
3 dataset (webvid-covr.zip), and a video (covr.mp4) as separate files.

4 A Dataset statistics

5 As explained in Section 3.1 of the main paper, we filter caption pairs with CLIP text embedding
6 similarity ≥ 0.96 and caption pairs with CLIP text embedding similarity ≤ 0.6 , and for each caption
7 pair, we choose the 10 video pairs with the highest CLIP visual similarity computed at the middle
8 frame of the videos. We also note that our cosine similarities are normalized between $[0, 1]$. Here,
9 we further show the distribution of text embedding similarity in caption pairs and visual embedding
10 similarity in video pairs in Figure A.1. The distribution of video similarity scores exhibits two distinct
11 peaks. The first peak corresponds to a score of approximately 0.7 and includes video pairs that are
12 significantly dissimilar. The second peak corresponds to a score close to 1.0 and represents video
13 pairs with highly similar visual content.

14 Figure A.2 further provides the histogram of the number of words in the generated modification text.
15 We observe that the majority of texts contain 3-8 words.

16 In Section 3.2 of the main paper, we provided several statistics about our WebVid-CoVR dataset, e.g.,
17 on average, a target video is associated with 12.7 triplets. However, in Figure A.3, when visualizing
18 the distribution of triplets associated with each target video, we see that the histogram reveals that
19 the majority of target videos are associated to only 1 or 2 triplets. The histogram exhibits a long tail,
20 i.e., a small subset of target videos have a considerably larger number of triplets associated. These
21 videos have captions such as “Mountain landscape”, “Water stream”, and “Water river”, leading to
22 numerous one-word difference captions associated with them.

23 B Implementation details

24 We describe further training details (Section B.1), provide the templates we use for our rule-based
25 baseline (Section B.2), and details about our MTG-LLM finetuning and inference (Section B.3).

26 B.1 Training details

27 Here, we provide implementation details in addition to Section 4.1 of the main paper of the main
28 paper. In terms of the optimization algorithm, we utilize AdamW [2]. For our MTG-LLM, we
29 finetune for one epoch with a batch size of 128 and a learning of $3e^{-5}$ that is warmed up linearly
30 for the first 100 steps and then kept constant. For our CoVR model, keeping the visual backbone
31 frozen largely improves the efficiency of the training process: an epoch on the CIRRR dataset takes 4
32 minutes with a frozen backbone and 25 minutes with a finetuned backbone, while leading to similar

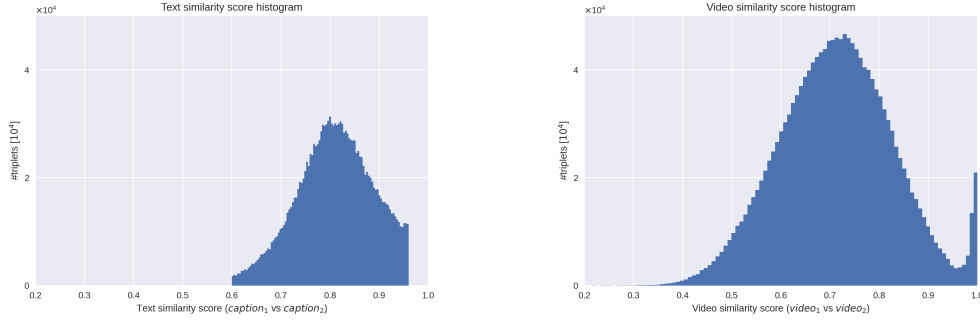


Figure A.1: **Text/video similarity of the caption/video pairs:** Distribution of text similarity scores between caption pairs ($caption_1, caption_2$) (left) and video similarity scores between video pairs ($video_1, video_2$) (right), using CLIP embeddings and cosine similarity.

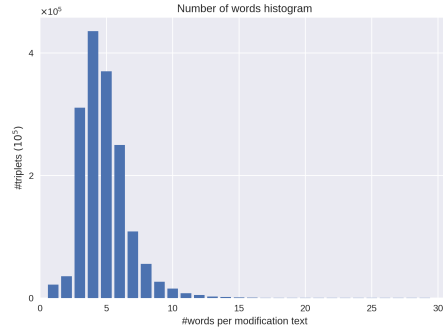


Figure A.2: **Histogram of the number of words in the generated modification text:** Most modification texts have between 3 and 8 words.

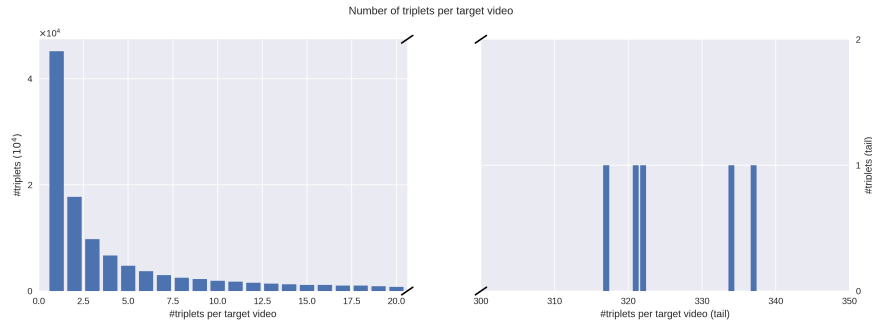


Figure A.3: **Distribution of number of triplets per target video:** We display the histogram depicting the number of triplets associated with each target video in the WebVid-CoVR dataset. Most target videos have 1 or 2 triplets and certain videos exhibit a high number of triplets (zoomed in to the tail on the right plot), e.g., some target videos are present in over 300 triplets, highlighting the variability in modification texts.

Table A.1: **Rule-based templates:** For our rule-based MTG baseline, we randomly choose one of the below templates during training.

Remove txt_diff₁
Take out txt_diff₁ and add txt_diff₂
Change txt_diff₁ for txt_diff₂
Replace txt_diff₁ with txt_diff₂
Replace txt_diff₁ by txt_diff₂
Replace txt_diff₁ with txt_diff₂
Make the txt_diff₁ into txt_diff₂
Add txt_diff₂
Change it to txt_diff₂

performance. During the training process, we employ several image data augmentations. These transformations include a random resized crop, where the input image is resized to a resolution of 384×384 . Additionally, we apply a random horizontal flip and random adjustments to contrast, brightness, sharpness, translation, and rotation. We use a weight decay of 0.05 and an initial learning rate of $1e^{-5}$ that is decayed to 0 following a cosine schedule over 10 epochs.

B.2 List of rule-based templates

In the ablation studies (Section 4.2 of the main paper), we introduced a rule-based MTG baseline. Here, in Table A.1, we show the templates used for the rules. We refer to Section D.2 (Table A.5) for qualitative comparison with our finetuned MTG-LLM.

B.3 Generating a modification text from paired captions with MTG-LLM

As described in Section 3.1, we use top-k sampling at inference for the MTG-LLM. Specifically, we use $k = 200$ and $temperature = 0.8$. We further give details about the text input-output format for the MTG-LLM. At training, we form the input prompt by concatenating captions and target and adding delimiters and stop sequences similar to InstructPix2Pix [1]. In detail, given a caption pair $(caption_1, caption_2)$ and a corresponding target $Target$, we concatenate them and add a separator in the following way: $caption_1\{\text{separator}\}caption_2\backslash n\&\&\backslash nTarget$, where separator is $\backslash n\&\&\backslash n$.

For instance, the model takes as input:

Clouds in the sky\&\&\nAirplane in the sky \n\n### Response :

and is trained to generate the response:

Clouds in the sky\&\&\nAirplane in the sky \n\n### Response :
Add an airplane

At inference, we simply leave the response empty, and let the model autoregressively generate a modification text.

As mentioned in Section 3.1 of the main paper, we add 15 manually prepared text triplets to the existing 700 text triplets from [1] used for training. The motivation is to address specific CoVR cases not present in the original set of triplets, such as “remove clouds and reveal only sky” given input captions “Clouds timelapse” and “Sky timelapse”. We show these 15 samples in Table A.2.

C Additional experiments

We provide additional experiments, reporting CoVR results obtained by training on training data generated with prompting (i.e., without finetuning) the LLM (Section C.1), and with varying training batch size (Section C.2).

Table A.2: **Added examples to the MTG-LLM training:** We add the below 15 examples to the set of 700 text triplets from [1].

Caption ₁	Clouds in the sky
Caption ₂	Airplane in the sky
Target output	Add an airplane
Caption ₁	Woman with the tablet computer sitting in the city.
Caption ₂	Woman with tablet computer sitting in the park.
Target output	In the park
Caption ₁	Walking swan
Caption ₂	White swan
Target output	Change color to white
Caption ₁	Child playing on beach, sea waves view, girl spinning on coastline in summer 4k
Caption ₂	Child playing on beach, sea waves view, girl running on coastline in summer 4k
Target output	Make her spin
Caption ₁	Aerial view of forest
Caption ₂	Aerial view autumn forest
Target output	Change season to autumn
Caption ₁	Palm tree in the wind
Caption ₂	Palm trees in the wind
Target output	Add more palm trees
Caption ₁	Schoolgirl talking on the phone
Caption ₂	Girl talking on the phone
Target output	Make her older
Caption ₁	Clouds timelapse
Caption ₂	Sky timelapse
Target output	remove clouds and reveal only sky
Caption ₁	Aerial view of a sailboat anchored in the mediterranean sea, vathi, greece.
Caption ₂	Aerial view of two sailboat anchored in the mediterranean sea, vathi, greece.
Target output	Add one sailboat
Caption ₁	France flag waving in the wind. realistic flag background. looped animation background.
Caption ₂	Italian flag waving in the wind. realistic flag background. looped animation background.
Target output	Swap the flag for an italian one
Caption ₁	Woman jogging with her dog in the park
Caption ₂	Woman playing with her dog in the park.
Target output	Stop jogging and make them play
Caption ₁	Oil Painting Reproductions of by humans william-glackens
Caption ₂	Oil Painting Reproductions of zombies by william-glackens
Target output	Replace the humans with zombies
Caption ₁	The girl who loved the sea by banafria
Caption ₂	The girl, wearing a hat, who loved the sea by banafria
Target output	Put a hat on her
Caption ₁	famous painting Paris, a Rainy Day of Gustave Caillebotte
Caption ₂	famous painting Paris, a Sunny Day of Gustave Caillebotte
Target output	Change it to more pleasant weather
Caption ₁	Bee on purple flower
Caption ₂	Bee on a flower
Target output	Change color of the flower

Table A.3: **Prompting versus finetuning LLM:** We compare our finetuned model (MTG-LLM) to a prompting baseline (see Section C.1) and observe important gains in the downstream performance of the model trained on the generated data.

Model	WebVid-CoVR _m				CIRR			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
Prompting	52.98	78.23	85.87	97.17	34.75	62.94	73.95	89.81
Finetuning	54.87	80.99	88.30	98.11	38.55	66.80	77.25	91.61

Table A.4: **Batch size:** We report results when training with two different batch sizes. We observe similar performance when reducing the batch size to 1024.

	WebVid-CoVR _m				CIRR			
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50
1024	54.91	80.90	87.97	98.23	38.43	66.70	76.87	91.57
2048	54.87	80.99	88.30	98.11	38.55	66.80	77.25	91.61

64 C.1 Prompting versus finetuning the MTG-LLM

65 Here, we justify why we finetuned Llama as opposed to simply prompting it without any training. For
66 prompting, we prepend few-shot examples of pairs of captions and desired generated texts, before
67 adding the two captions in question. In particular, we use the following sentence:

68 Clouds in the sky&&Airplane in the sky-> Add an airplane\n
69 Aerial view of forest&&Aerial view autumn forest-> Change
70 season to autumn\n
71 Clouds timelapse&&Sky timelapse-> remove clouds and reveal
72 only sky\n
73 Aerial view of a sailboat anchored in the mediterranean sea.&&
74 Aerial view of two sailboat anchored in the mediterranean
75 sea-> Add one sailboat\n

76 Then, we concatenate our two captions for which we wish to generate a modification text. Table A.3
77 shows that finetuning the MTG-LLM for generating the training data is much more effective than
78 prompting it without finetuning, as measured by CoVR performance on WebVid-CoVR_m and CoIR
79 performance on CIRR. This is also consistent with our qualitative observations: we found that the
80 LLM struggles to perform the modification text generation without finetuning (see Table A.5 in the
81 next section).

82 C.2 Batch size

83 As mentioned in Section 4.1 of the main paper, our batch size throughout experiments is 2048.
84 In Table A.4, we experiment whether this hyperparameter influences the results. We observe that
85 decreasing the batch size to 1024 does not have a significant impact on the performance on WebVid-
86 CoVR_m and CIRR.

87 D Qualitative analysis

88 In this section, we provide examples of caption filtering (Section D.1), qualitative comparison
89 between different MTG approaches (Section D.2), qualitative examples of our WebVid-CoVR triplets
90 (Section D.3), samples from our manual test set annotation process (Section D.4), qualitative CoVR
91 results on WebVid-CoVR_m (Section D.5) and CoIR results on CIRR (Section D.6).

92 D.1 Examples of filtered captions

93 As described in Section 3.1 of the main paper, we employ a filtering process to select paired captions
94 that facilitate the generation of meaningful training data. In this section, we provide examples of the
95 filtered captions.

96 **Filtering template captions.** Upon analyzing the paired captions, we observed that a significant
 97 portion of the pairs originated from a small set of template captions. Out of 1.2M distinct caption pairs,
 98 approximately 719k (60%) were generated from these template captions. The following examples
 99 showcase some of these template captions:

- 100 • **Abstract:** *Abstract color movement tunnel, Abstract color nature background, Abstract*
 101 *color smoke flowing on white background, Abstract colorful paint ink spread explode,*
 102 *Abstract colorful pattern background, Abstract colorful red cement wall background or*
 103 *texture. the camera moves up, Abstract colorful satin background animation, Abstract*
 104 *colorful shiny bokeh background., Abstract colorful smoke on black background, etc*
- 105 • **Background:** *Abstract background, Animated backgrounds, Animation, background., Aquar-*
 106 *ium background, Artistic background, Aurora background, Balloons background, Basket-*
 107 *balls background, Beach background, Bluebell background, Bright background, Brush*
 108 *background, Bubbles background, Bubbly background, Celebrate background, celebra-*
 109 *tory background, Cg background, Christmas background, Christmas background, Circles*
 110 *background, Color background, Colored background, Colorful background, Colorfull back-*
 111 *ground., etc.*
- 112 • **Concept:** *Brazil high resolution default concept, Brazil high resolution dollars concept,*
 113 *Businessman with advertising hologram concept, Businessman with algorithm hologram*
 114 *concept, Businessman with automation hologram concept, Businessman with bitcoin holo-*
 115 *gram concept, Businessman with branding hologram concept, Businessman with public*
 116 *relations hologram concept, Close up of an eye focusing on a freelance concept on a futuris-*
 117 *tic screen., Coins fall into piggy bank painted with flag of ghana. national banking system or*
 118 *savings related conceptual 3d animation, Communication concept, Communication network*
 119 *concept., Communication team concept, Concept of connection, Concept of dancing at disco*
 120 *party. having fun with friends., Concept of education, Concept of geography, Cyber monday*
 121 *concept, etc*
- 122 • **Flag:** *Flag of america, Flag of andorra, Flag of aruba, Flag of austria, Flag of azerbaijan,*
 123 *Flag of bahrain, Flag of belarus, Flag of belize, Flag of black, Flag of bolivia, Flag of brazil,*
 124 *Flag of bulgaria, Flag of cameroon, Flag of canada, etc.*

125 **Filtering caption pairs with high or low similarity.** To ensure the generation of meaningful
 126 modifications, we further refine the selection of caption pairs by filtering out those with excessively
 127 high or low similarity. Caption pairs with highly similar meanings may result in trivial or unnoticeable
 128 modifications. Conversely, pairs with significant dissimilarity can lead to large visual differences that
 129 are difficult to describe accurately. We show below some of the filtered captions based on the CLIP
 130 text embedding cosine similarity.

- 131 • **High similarity:** 10% of the pairs have CLIP text similarity above 0.96.
 - 132 – Close-up of a tree with green leaves and sunlight
 - 133 – Close-up of a tree with green leaves and sunshine
 - 134 – Businessman speaking on the phone
 - 135 – Businessman talking on the phone
 - 136 – Boat on a sea
 - 137 – Boat on the sea
- 138 • **Low similarity:** 2% of the pairs have CLIP text similarity below 0.60.
 - 139 – Leaves close-up
 - 140 – Peacock, close-up
 - 141 – Moon jellyfish
 - 142 – Moon night
 - 143 – Close up of a lynx
 - 144 – Close up of a milkshake

145 **Exclusion of digit differences and out-of-vocabulary words.** In order to maintain the high
 146 quality and coherence of the generated modification text, we apply additional filtering criteria.

Specifically, we exclude caption pairs where the differences between captions are numerical digits (often representing dates) or involve out-of-vocabulary words (using the python libraries wordfreq and enchant) that may hinder the generation process.

- **Difference between the captions is a digit:** Approximately 2% of the pairs.
 - 23.09.2015 navigation on the moscow river
 - 07.08.2015 navigation on the moscow river.
 - Light leaks element 190
 - Light leaks element 215
 - Pure silver, shape of granules of pure silver each one is unique 44 (2)
 - Pure silver, shape of granules of pure silver each one is unique 95 (2)
- **Difference in one of the captions has an out-of-vocabulary word:** Approximately 7% of the pairs.
 - Businessman writing on hologram desk tech word- bitcoin
 - Businessman writing on hologram desk tech word- crm
 - Mitomycin-c - male doctor with mobile phone opens and touches hologram active ingrident of medicine
 - Oxazepam - male doctor with mobile phone opens and touches hologram active ingri- dent of medicine
 - Blue forget-me-nots
 - Blue galaxy

D.2 Qualitative comparison of MTG approaches

In Section 4.4 of the main paper and Section C.1, we show that finetuning our MTG-LLM works better than a rule-based approach and than few-shot prompting of the LLM. In this section, we provide a qualitative comparison of three different methods for generating modification text: (i) rule-based, (ii) prompting-based, and (iii) our MTG-LLM finetuning. We present examples of paired captions and the corresponding modification texts generated by each method in Table A.5.

Rule-based method. The rule-based method relies on predefined rules to generate modification text. We illustrate an example limitation in the last row of Table A.5, where the difference text is simply a preposition (i.e., ‘of’ vs ‘above’), and the modification text becomes ‘Remove of’. The rule-based method performs well when the modifications follow a specific pattern, but it may struggle with more complex modifications (e.g., ‘tree’ vs ‘trees’ should generate ‘add more trees’ for plurality).

Prompting LLM. The prompting-based method involves using a pretrained language model without finetuning. However, this method is prone to hallucinations and may generate modification text that does not accurately represent the intended difference. For example, in the second example, the prompting LLM suggests removing the term ‘animal’ instead of replacing ‘bird’ with ‘bear’.

MTG-LLM (Our approach). Our MTG-LLM approach utilizes a large language model finetuned on a manually annotated dataset specifically for modification text generation. It tends to be the most robust across different cases.

D.3 Training triplet examples

Figures A.4, A.5, and A.6 all show examples of triplets generated using our automatic dataset creation. These examples demonstrate the effectiveness of our approach in generating coherent modification texts for paired videos. This capability serves as a form of data augmentation and increasing the diversity in the training set. In Figure A.7, we show that the dataset is not composed by pairs only, as there are many captions that have many relations between them. Furthermore, in Figure A.8 we show cases where a single caption is associated with multiple videos. This scenario allows us to generate multiple triplets by leveraging the diverse visual content captured in different videos. The triplets shown in the aforementioned figures exhibit a wide range of variations, encompassing different themes such as emotions, food, actions, camera edits, gender changes, and time of the day.

Table A.5: **Comparison between modification text generation approaches:** We provide qualitative examples for a pair of captions, and three methods to generate modification text: (i) rule-based, (ii) prompting-based, (iii) our MTG-LLM finetuning. Rule-based method is limited, for example in the case where the difference text is a preposition (last row), whereas the prompting-based method is prone to hallucinating (e.g., ‘remove iceberg’, ‘change the pose of the runner’). Our approach tends to be the most robust across cases.

Caption ₁	<i>Happy</i> girl dancing
Caption ₂	<i>Beautiful</i> girl dancing
Rule-based	Make the <i>Happy</i> into <i>Beautiful</i>
Prompting LLM	Change girl
MTG-LLM	Make her more <i>beautiful</i>
Caption ₁	Black <i>bird</i>
Caption ₂	black <i>bear</i>
Rule-based	Add <i>bear</i>
Prompting LLM	Remove animal
MTG-LLM	Turn it into a <i>bear</i>
Caption ₁	<i>Autumn</i> landscape in the mountains.
Caption ₂	<i>Winter</i> landscape in the mountains
Rule-based	Remove <i>Autumn</i>
Prompting LLM	Change scene to winter.
MTG-LLM	Change to <i>winter</i>
Caption ₁	Flying over the <i>ice</i>
Caption ₂	Flying over the <i>mountains</i>
Rule-based	Take out <i>ice</i> and add <i>mountains</i>
Prompting LLM	Remove iceberg
MTG-LLM	Move the <i>ice</i> to the <i>mountains</i>
Caption ₁	<i>Dandelion</i> field
Caption ₂	<i>Rice</i> field
Rule-based	Change <i>dandelion</i> for <i>rice</i>
Prompting LLM	Change the season
MTG-LLM	Turn it into a <i>rice</i> field.
Caption ₁	<i>Happy</i> woman
Caption ₂	<i>Running</i> woman
Rule-based	Change it to <i>running</i>
Prompting LLM	Remove the girl and change the pose of the runner
MTG-LLM	Make her a runner
Caption ₁	Woman working on cellphone at <i>night</i>
Caption ₂	Woman working on cellphone at <i>outdoor</i>
Rule-based	Replace <i>night</i> with <i>outdoor</i>
Prompting LLM	Change the woman and add one cellphone
MTG-LLM	Make it an <i>outdoor</i>
Caption ₁	Barber cuts the hair of the client with <i>clipper</i> .
Caption ₂	Barber cuts the hair of the client with <i>trimmer</i>
Rule-based	Take out <i>clipper</i> and add <i>trimmer</i>
Prompting LLM	Add or change the hairdresser’s tool
MTG-LLM	Change to a <i>trimmer</i>
Caption ₁	Aerial shot <i>of</i> a lake.
Caption ₂	Aerial shot <i>above</i> a lake
Rule-based	Remove <i>of</i>
Prompting LLM	Add a lake to the sequence
MTG-LLM	make it an aerial shot <i>of</i> a lake

195 D.4 Manual test set annotation

196 In this section, we further describe the process of manually annotating the test set for our WebVid-
197 CoVR_m CoVR benchmark, previously discussed in Section 3.2. The annotation process involves
198 presenting the annotator with generated modification texts from three different runs of MTG-LLM,
199 along with three frames each from the query and target videos. The annotator’s task is to evaluate the
200 quality of the modification texts and the suitability of the videos for the CoVR task.

201 A total of 3.1K triplets were shown for annotation. In Figure A.9 and Figure A.10, we present 10
202 examples that were considered correct during the annotation, along with the chosen modification
203 texts (marked with a checkmark). These examples demonstrate successful modification texts and
204 appropriate video content for the CoVR task.

205 On the other hand, in Figure A.10, we show 8 examples that were discarded during the annotation.
206 These examples were rejected either because the modification texts were incorrect or because the
207 videos were deemed unsuitable for the CoVR task due to being either too similar (e.g., bottom left,
208 both videos are showing the same coffee with almost no modification) or too incoherent (e.g., top right
209 example “Make the water a river”).

210 D.5 Qualitative CoVR results on WebVid-CoVR_m

211 In Figure A.11, we show qualitative CoVR results on our manually verified WebVid-CoVR_m test set.
212 We observe that top ranked video frames have high visual and semantic similarity with the queries
213 even when not corresponding to the ground truth (marked with a green border).

214 D.6 Qualitative CoIR results on the CIRRR benchmark

215 In Figure A.12, we demonstrate qualitative CoIR results of our models trained only on WebVid-CoVR
216 (ZS) and the one further finetuned on CIRRR training set (Sup.), tested on the CIRRR test set. We
217 observe promising retrieval quality for both models.

218 References

- 219 [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow
220 image editing instructions. *arXiv:2211.09800*, 2022. 3, 4
- 221 [2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1



Figure A.4: **Examples of generated triplets:** We illustrate triplet samples (one per row) generated using our automatic dataset creation methodology. Each sample consists of two videos with their corresponding captions (at the bottom of each video) and the generated modification text using our MTG-LLM (in purple).

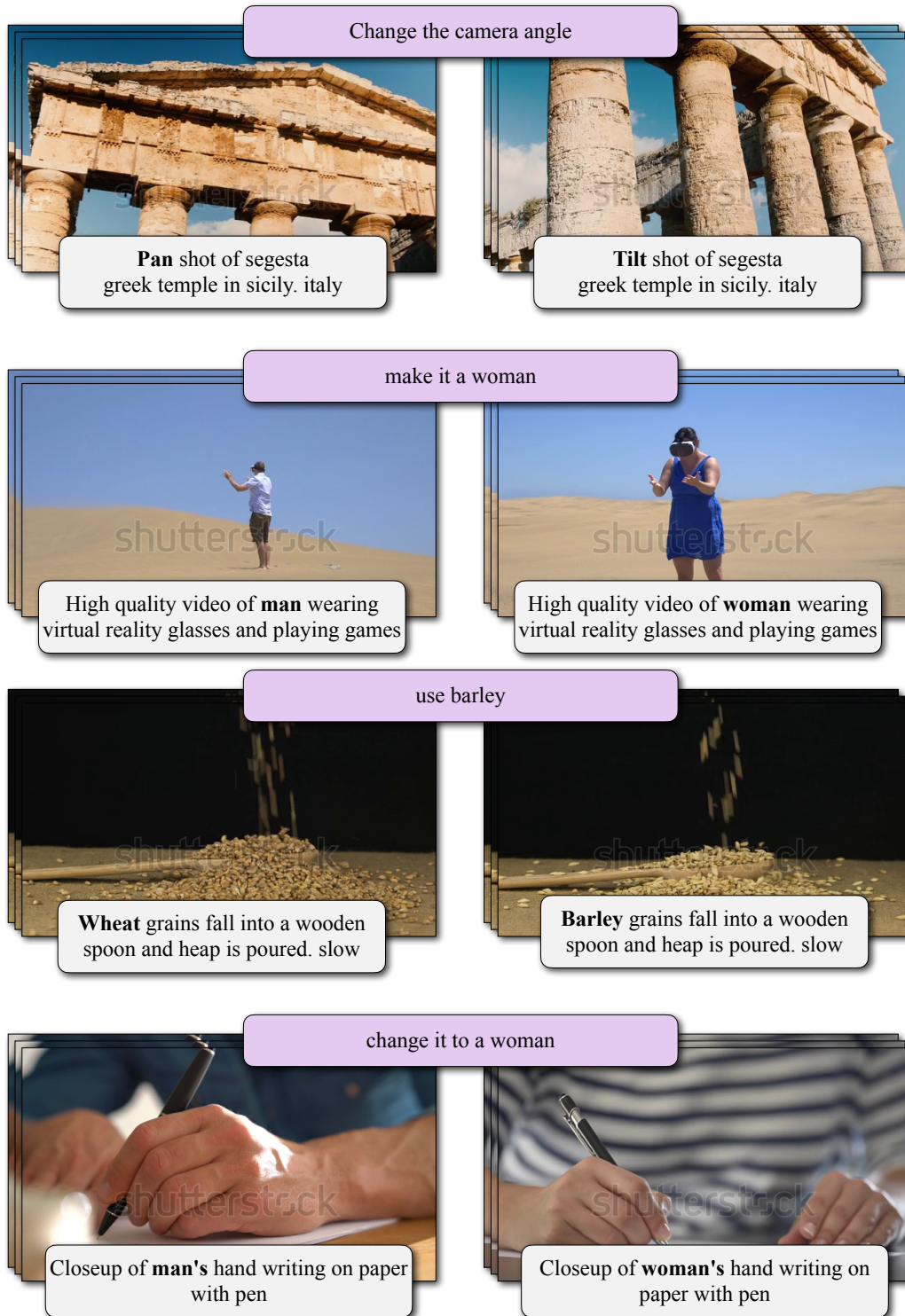


Figure A.5: Examples of generated triplets (ctd)

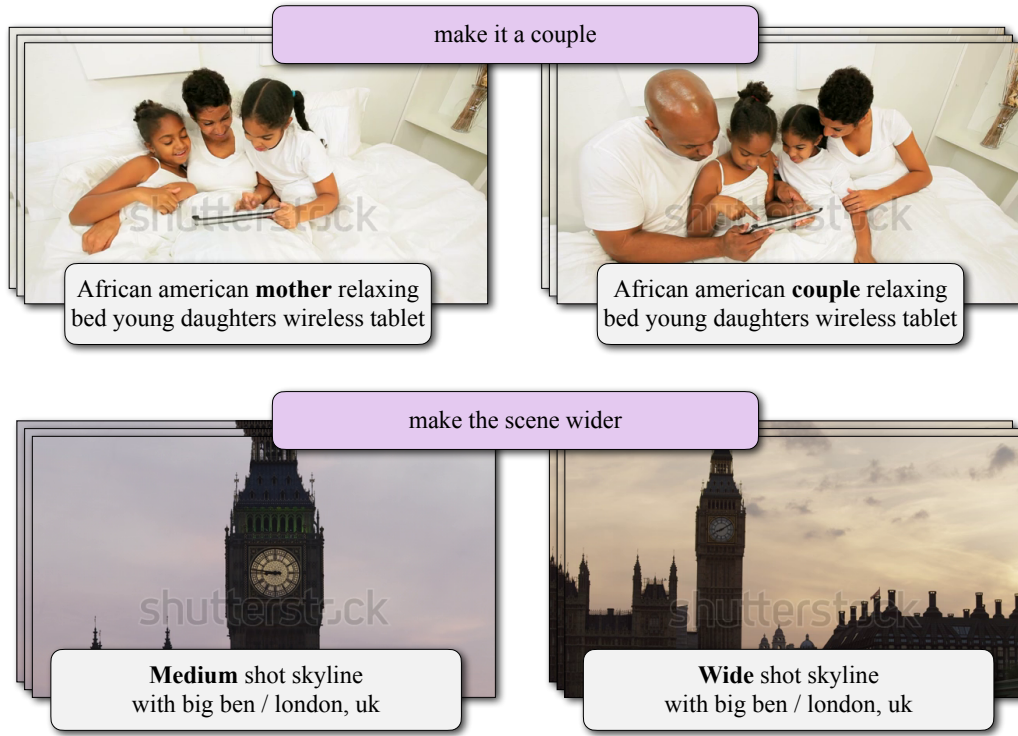


Figure A.6: Examples of generated triplets (ctd)

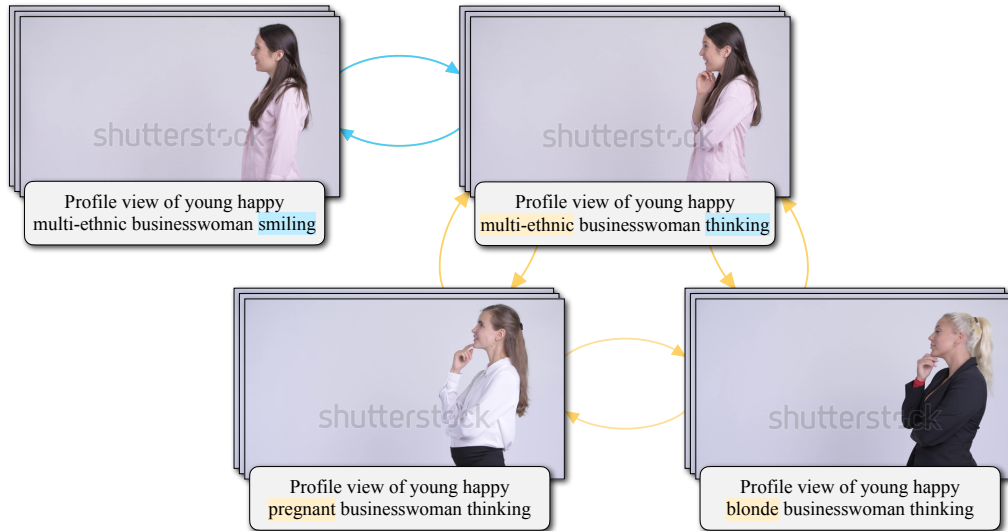


Figure A.7: **Generated triplets from multiple similar captions:** We can train with as many triplets as pairs of captions with one word difference by generating modification texts using our trained MTG-LLM: *she is thinking* , *Have her look happy* , *Make the businesswoman pregnant* , *make her blonde* , *make her multi-ethnic* , *Make the woman pregnant* , etc.



Figure A.8: **Generated triplets with multiple videos:** In cases where there are several videos with the same caption, we can generate multiple triplets by leveraging the multiple videos. It can be seen as a way of data augmentation.

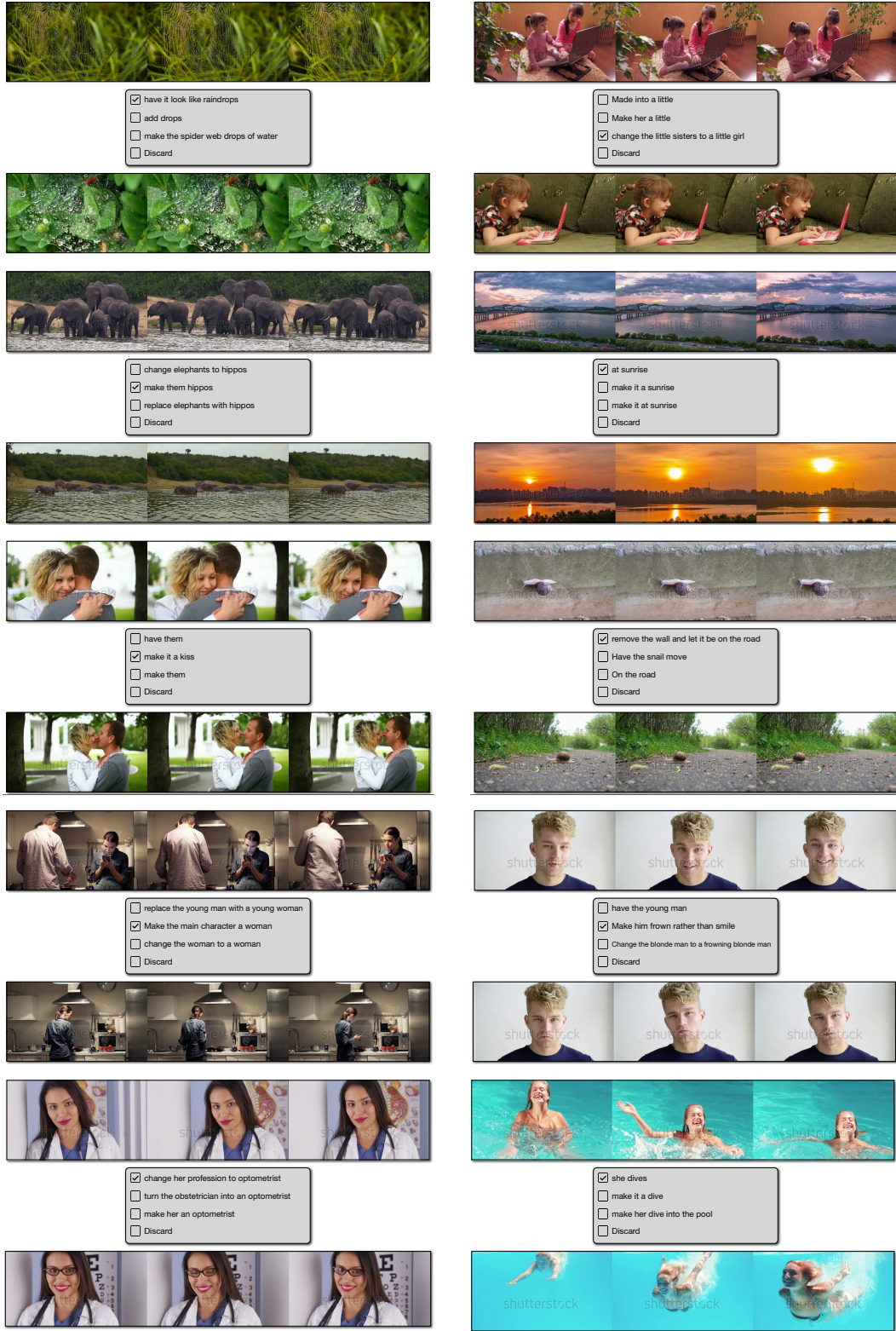


Figure A.9: **Manual annotation examples (kept):** We show samples from WebVid-CoVR_m which are automatically mined triplets that are marked as correct during the annotation process. Each sample consists of two videos and a set of modification text options (in between each video pair). The chosen modification text is indicated by a checkmark.

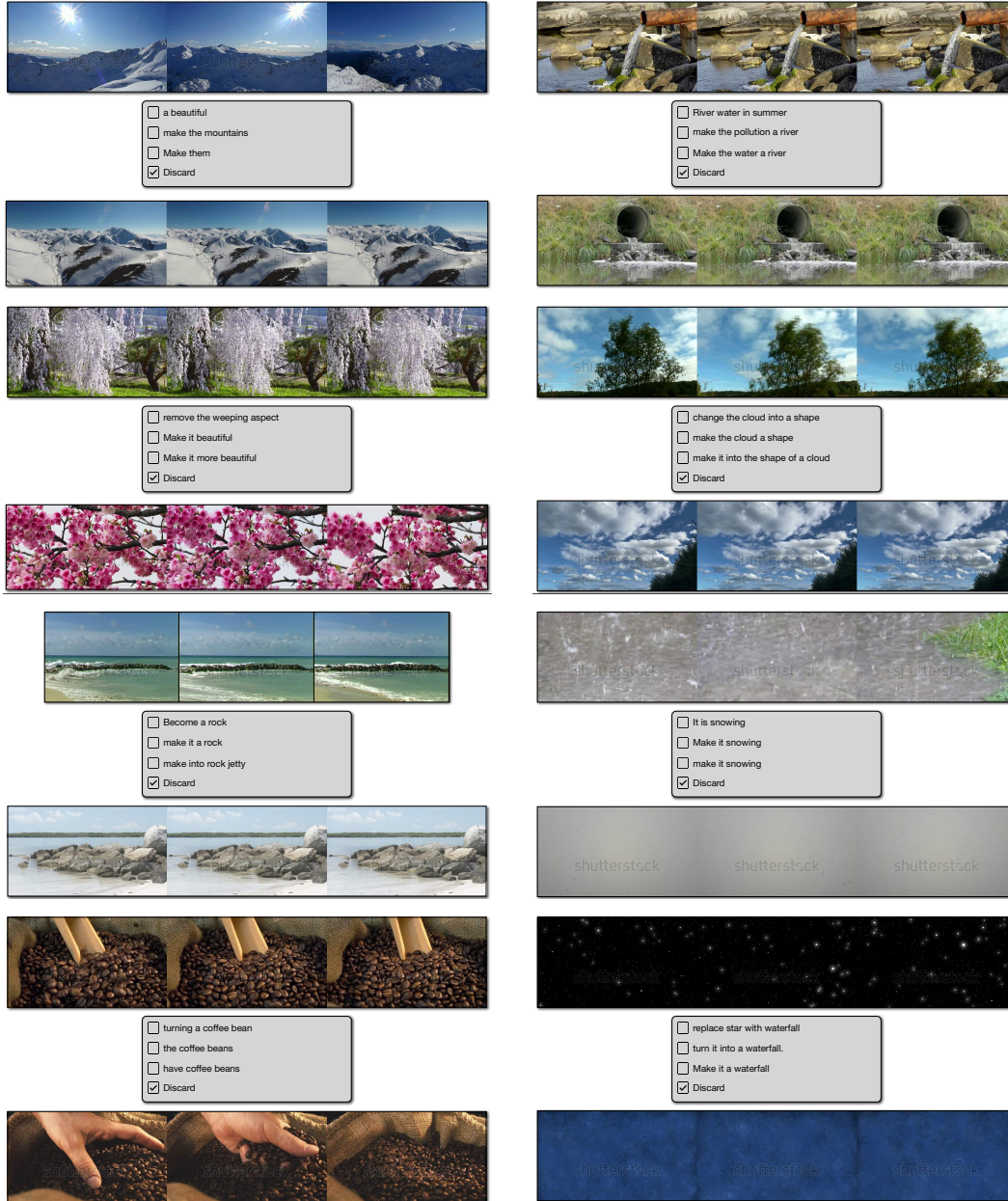


Figure A.10: **Manual annotation examples (discarded):** We show automatically mined triplets that are discarded during the annotation process. Discarded texts include videos that are too similar (bottom left), too dissimilar (bottom right), or have bad modification texts (top left).

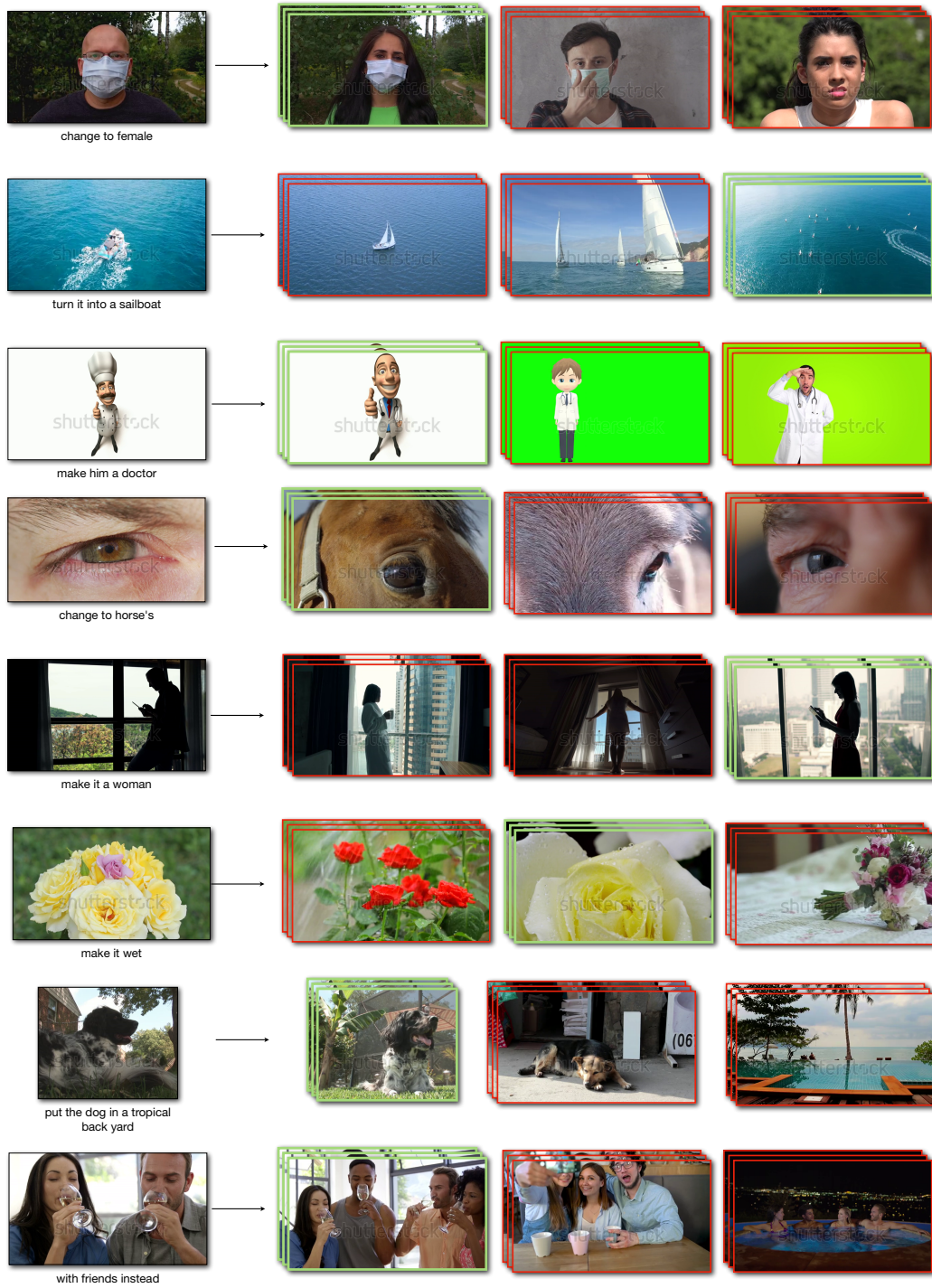


Figure A.11: **Qualitative CoVR results on WebVid-CoVR_m**: We display the input image and modification text queries on the left, along with the top 3 retrieved videos by our model on the right. Ground-truth is denoted with a green border.

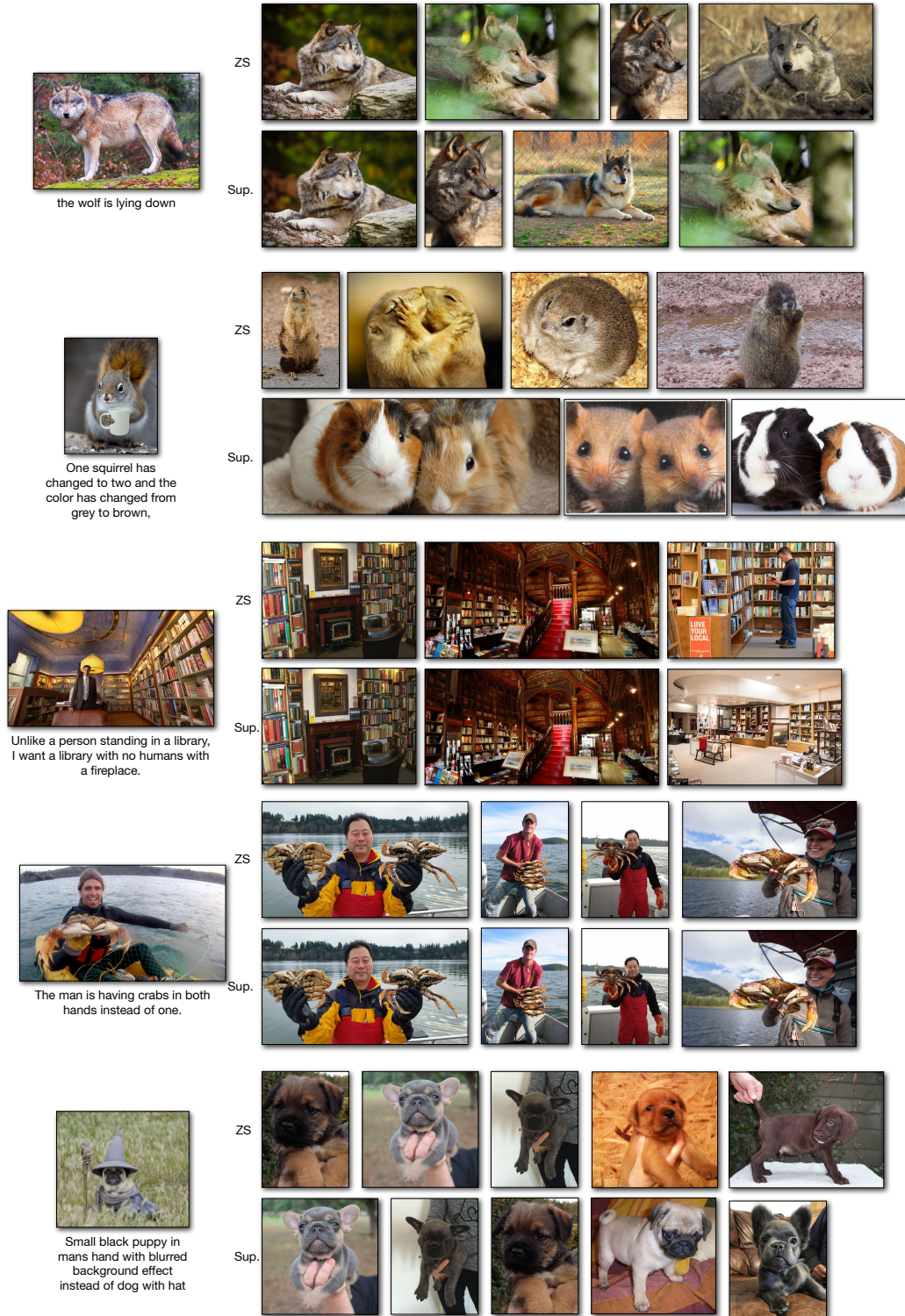


Figure A.12: **Qualitative CoIR results on CIRR:** Given a query image and a modification text, we show our top retrieved videos of our zero-shot (ZS) model trained with WebVid-CoVR and the model finetuned on CIRR ground-truth supervision (Sup.).