

Appendix

A. Additional Editing Examples

We provide additional qualitative examples of the editing results on the TEdBench benchmark, generated based on the experimental setup detailed in Section 5.2. These examples further supplement those presented in Figures 1 and 4. Each example is accompanied by the temporal editing caption used to perform the edit, which was generated using the method described in Section 3.1. Furthermore, in Figure S9, we supplement Figure 1 with additional video examples generated by our method, illustrating the transition from the source image (left) to the target edit (right).









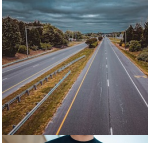

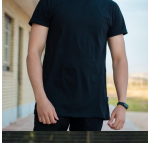



Source Image	Target Editing Caption	Temporal Editing Caption	Edited Image
	A photo of a vase of red roses.	The carnations slowly transform into vibrant red roses.	
	Pizza with pepperoni.	Pepperoni slices are carefully placed on the pizza.	
	A photo of a couple holding their hands on a beach.	The couple on the beach gradually moves closer to each other and holds hands.	
	A photo of a tree. Taken in autumn.	The tree's green leaves gradually turn vibrant shades of red and orange.	
	A photo of a traffic jam.	Cars begin to fill the once-clear highway, gradually slowing to a standstill.	
	A person in a greeting pose to Namaste hands.	The person's hands slowly move together into a Namaste position.	
	A photo of a sitting dog.	The dog gradually lowers its back legs, settling into a seated position.	

Figure S7. TEdBench Editing Examples.



















Source Image	Target Editing Caption	Temporal Editing Caption	Edited Image
	A goat jumping over a cat.	The goat leaps gracefully into the air over the unmoving cat.	
	A photo of a car in Manhattan.	The car drives through a bustling Manhattan street.	
	A girl riding a horse.	The horse lowers its head, revealing a girl seated on its back.	
	A teddy bear holding a cup.	The teddy bear gently lifts a cup into its arm.	
	A photo of an open box.	The flaps of the box gently unfold, revealing the open interior.	
	A photo of a giraffe eating the grass below.	The giraffe gently lowers its neck to nibble on the grass.	
	Two bananas.	A second banana gently rolls into view, coming to rest beside the first banana.	
	A cyclist riding in a street.	A cyclist gradually emerges, pedaling along the cobblestone street.	
	A white horse in a grass field.	The snow beneath the white horse melts, revealing a lush grass field.	

Figure S8. TEdBench Editing Examples.

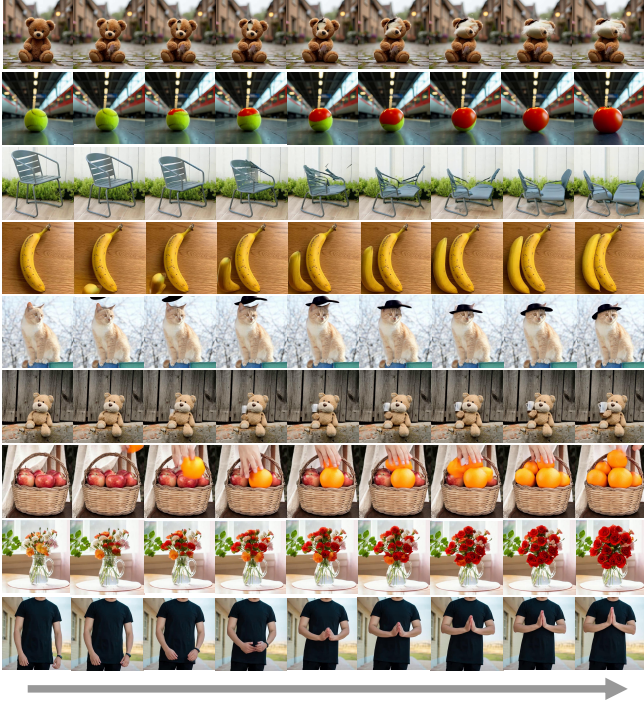


Figure S9. **Generated Video Examples.** Additional video sequences illustrating the temporal evolution from the source image to the target edit.

B. Temporal Editing Captions

B.1. VLM Instruction

As outlined in Section 3.1, we propose a framework for automatically generating the temporal editing caption by leveraging the original target editing prompt in conjunction with the source image. The instruction given to the VLM, along with the source image is:

‘Write a one-sentence description of a short video that begins with the provided image and smoothly transitions into a scene of a “CAPTION”, highlighting how elements in the image undergo changes or movement over time. Keep the description simple, concise and short, focusing only on essential changes and actions without altering unnecessary details. Avoid mentioning elements that do not contribute to the main change needed, and focus the description on the main transitions. Do not add objects that are not in the original image or described in the final scene. The camera should remain static unless movement is absolutely necessary. Ensure all transitions happen within a few second duration without mentioning the length or using the word “video”.’

Here, “CAPTION” is replaced with the target caption specific to the image. Additionally, as explained, in-context learning is employed to provide the VLM with examples alongside the instruction. Before processing the desired

source image and edit prompt, the instruction is presented to the VLM nine times, each paired with a distinct example consisting of a source image, target caption, and corresponding temporal editing caption. Examples of these are illustrated in Figure S10.




	Target Editing Caption	Temporal Editing Caption
	A photo of domes in the Grand Canyon under the golden sunrise.	The sun rises slowly in the early morning.
	Two red Campari shots	Red-colored Campari is poured into shot glasses until they are filled.
	A photo of a magician holding a hat with a rabbit in it.	A magician slowly pulls a rabbit from his hat and reveals it gradually.

Figure S10. **In Context Learning Examples.**

B.2. Ablation

To assess the impact of the Temporal Editing Caption, we conduct an ablation experiment comparing its use against directly using the target editing captions from the TEd-Bench benchmark. Apart from this modification, we adhere to the same protocols as described in the original experiment in Section 5.2. As shown in Table S4, this setup preserves a similar resemblance to the source image but underperforms in terms of image editing performance.

Model	Source LPIPS _↓	CLIP-I _↑	Target CLIP _↑
Original Captions	0.21	0.89	0.60
Temporal Captions	0.22	0.89	0.63

Table S4. **Temporal Editing Captions Ablation.**

C. Frame Selection

C.1. VLM Instruction

As detailed in Section 3.3, our method selects the frame that best aligns with the intended edit from each generated video. To automate this process, inspired by [23], we create a collage of uniformly sampled frames from the video, along with the source image and target editing caption, and prompt a VLM to identify the optimal frame. The model is

instructed to select the earliest frame (i.e., with the lowest index) that satisfies the editing intent, minimizing deviation from the original image. In both this process and the best seed selection process (applied across all methods), if none of the edited frames successfully fulfill the desired edit, the original image is retained as the final output.

The instruction provided to the VLM is:

‘The image displays the source photo at the top, with a collage of 12 edited versions beneath it. The target edit image caption was: “CAPTION”. Your task is to choose the image from 1 to 12 that best follows this edit fully and naturally. If none of the images follows the edit, select image 0. If multiple images follow the edit equally, prioritize the one with the lowest number possible. Avoid selecting images that appear to follow the edit but are not edits of the original image. Additionally, avoid images where camera motion, zoom, or image quality differs significantly, or where the content does not appear stable relative to the original source. Respond with: “The selected edit is:x” where x is the number of your chosen edit.’

Here, “CAPTION” refers to the target editing caption. Examples of the collages can be found in Figure S11.

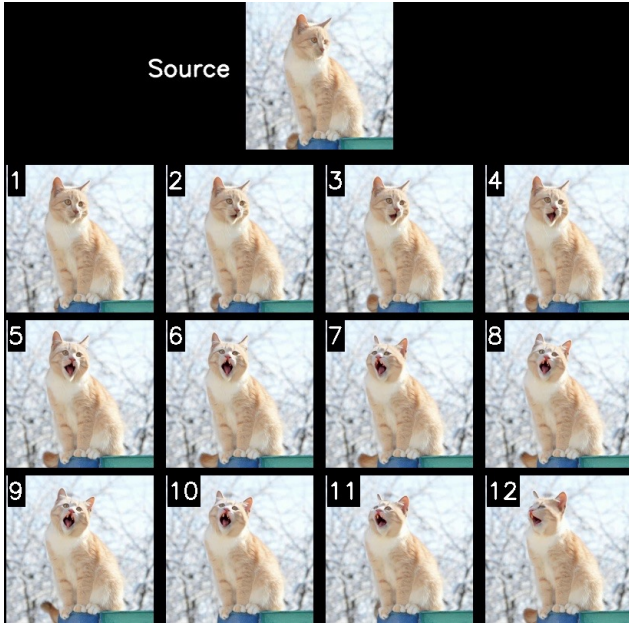


Figure S11. **Frame Selection Collage.** The target editing caption for this example is: “A photo of a cat yawning.”.

C.2. Ablation

To validate the effectiveness of our approach, we compare it to the naive solution of using the last frame of the generated video as the edited output. The evaluation follows the same protocol described in Section 5.2. As can be seen in Table S5, this naive approach results in a lower target CLIP

score for the edited outputs, highlighting the advantages of our method.

Model	Source		Target CLIP \uparrow
	LPIPS \downarrow	CLIP-I \uparrow	
Last Frame	0.24	0.9	0.61
Selected Frame	0.22	0.89	0.63

Table S5. **Frame Selection Ablation.**

D. Editing Manifold Pathway

As elaborated in Section 4, to simulate the image manifold, we generated 200 images across three distinct categories using FLUX.1-dev. Examples of these generated images are shown in Figure S12.



Figure S12. **Flux.1-dev Generations**

E. Inference Hyperparameters

We use the 5B-parameter image-to-video version of CogVideoX, ensuring consistency by applying the same inference hyperparameters across all experiments. The model generates a fixed 49 frames per sample. During inference, we use the default denoising scheduler in CogVideoX, which is based on DDIM with V-prediction. We perform 50 denoising steps and set the classifier-free guidance scale for text conditioning to 6.0.

F. PosEdit

In Section 5.3, we outline the construction of a human pose editing dataset. This dataset encompasses 58 editing tasks, distributed across 8 distinct action categories, featuring 8 different subjects. The source images consistently depict a neutral standing pose with arms relaxed at the sides, while the target poses vary according to the edit category. Each editing category is paired with a target caption and a temporal caption. Figure S13 and Figure S14 illustrates examples for each action category. Additionally, in Figure S15, we complement Figure 5 with quantitative results, demonstrating how the numerical evaluation aligns with its intended measurement objectives.

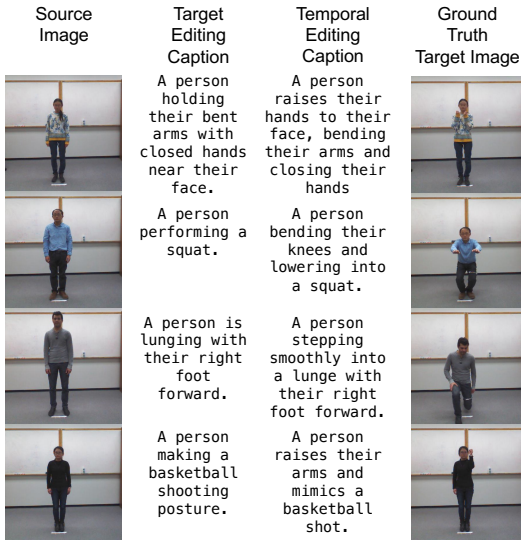


Figure S13. PosEdit Examples.

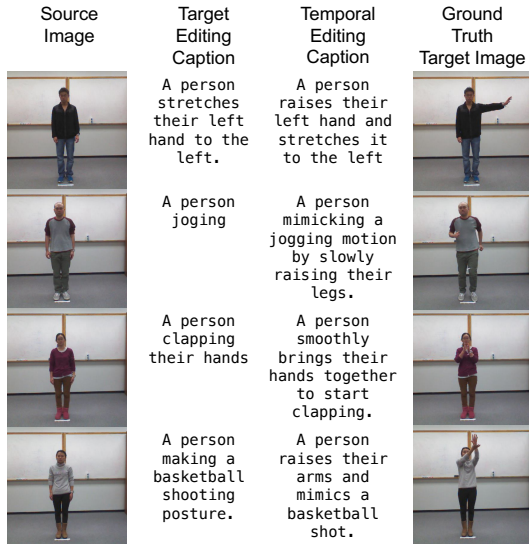


Figure S14. PosEdit Examples.

Source	LEDITS++	F2F	Ground Truth
"A person performing a squat."	Source LPIPS 0.28, CLIP-I 0.66 Target LPIPS 0.29, CLIP-I 0.71, CLIP 0.68	Source LPIPS 0.17, CLIP-I 0.78 Target LPIPS 0.14, CLIP-I 0.83, CLIP 0.70	Source LPIPS 0.10, CLIP-I 0.86 Target LPIPS -, CLIP-I -, CLIP 0.53
"A person jogging."	Source LPIPS 0.38, CLIP-I 0.65 Target LPIPS 0.39, CLIP-I 0.65, CLIP 0.62	Source LPIPS 0.13, CLIP-I 0.80 Target LPIPS 0.11, CLIP-I 0.81, CLIP 0.81	Source LPIPS 0.08, CLIP-I 0.88 Target LPIPS -, CLIP-I -, CLIP 0.51
"A person making a basketball shooting posture."	Source LPIPS 0.18, CLIP-I 0.59 Target LPIPS 0.19, CLIP-I 0.72, CLIP 0.44	Source LPIPS 0.13, CLIP-I 0.80 Target LPIPS 0.11, CLIP-I 0.81, CLIP 0.81	Source LPIPS 0.11, CLIP-I 0.88 Target LPIPS -, CLIP-I -, CLIP 0.79

Figure S15. PosEdit Quantitative Evaluation Examples.

G. Human Survey

As detailed in Section 5.4, we conducted a human evaluation survey to assess our method's performance based on real user preferences. Following the framework in [49], the survey questions evaluated (1) the accuracy of the edit relative to the prompt and (2) the quality of the edit, defined as the preservation of visual fidelity to the source image. Each participant reviewed 20 edits, comparing our method with LEDITS++. Examples of the pages shown to the evaluators are provided in Figures S16 and S17.



Compare the edit instruction with the actual changes made in the edited images. Select one edit image that most accurately/consistently implements the edit instruction.


- ☐ Image 1
- ☐ Image 2

Select one edited image that exhibits the best image quality. (Some aspects you may consider include the preservation of visual fidelity from the original image, seamless blending of edited elements with the original image, and the overall natural appearance of the modifications, etc.)

- ☐ Image 1
- ☐ Image 2

Figure S16. Survey Example.

A goat jumping over a cat.



Compare the edit instruction with the actual changes made in the edited images. Select one edit image that most accurately/consistently implements the edit instruction.

☐ Image 1

☐ Image 2

Select one edited image that exhibits the best image quality. (Some aspects you may consider include the preservation of visual fidelity from the original image, seamless blending of edited elements with the original image, and the overall natural appearance of the modifications, etc.)

☐ Image 1

☐ Image 2

Figure S17. Survey Example.

H. Further Ablations

We present additional ablation results demonstrating the ability of our method to handle two distinct editing challenges: (1) diverse backgrounds and (2) out-of-video-distribution edits.

Backgrounds. Similar to our image generation approach in Section 4, we generate eight images of two objects with varying backgrounds. Each image generation prompt follows one of the two templates below, where LOCATION is replaced with one of the following settings: colorful amusement park, sandy beach, magical fairy tale forest, futuristic cityscape, old library, snowy mountain peak, bustling train station, and quaint village square.

1. “A small brown teddy bear sitting in LOCATION.”
2. “A tennis ball lies in LOCATION.”

Out-of-video-distribution editing. As discussed in Section 6, one might assume that our method would fail to perform edits requiring temporal transformations that deviate significantly from typical real-world videos (the model’s training set). To show that this is not necessarily the case, we selected two unreal editing processes, using the following temporal captions:

1. “The teddy bear slowly rips, revealing stuffing coming out.”
2. “The tennis ball gradually transforms into a ripe red tomato.”

These edits depict unrealistic events that do not ordinarily occur in real-life footage. However, as shown in Fig. S18, our method successfully handles both the background variations and the out-of-distribution nature of these transformations. Moreover, the full temporal sequences in Fig. S9 demonstrate that these edits occur seamlessly without any external interaction (the teddy bear rips apart spontaneously, and the tennis ball morphs magically into a tomato).



Figure S18. Ablation Examples.

I. Additional Vision Tasks Captions

As outlined in Section 5.5 and demonstrated in Figure 6, we showcase our framework’s applicability for additional, more classic vision tasks that are not typically classified as image editing. For these tasks, we employ Runway Gen-3 as our video generator. Empirically, these tasks required longer and more descriptive captions. The temporal editing captions used for each task are as follows:

1. **Relighting:** ‘The scene’s lighting shifts gradually, changing to night. The sun is setting, and artificial lights replace it. The camera is static. Time-lapse. Cinematic.’
2. **Outpainting:** ‘The image expands, adding new surroundings seamlessly beyond the original frame.’
3. **Denoising:** ‘The image clears up as noise fades away, revealing smoother, cleaner details.’
4. **Deblurring:** ‘The camera comes into focus, revealing sharp details and enhanced clarity, as though a camera lens has adjusted perfectly. Nothing moves. Static image.’