Long Video Diffusion Generation with Segmented Cross-Attention and Content-Rich Video Data Curation

Appendix

A. More Related Work

Video Generation Datasets are crucial for pre-training high-quality video generation models. Existing text-video datasets [3] have made substantial progress in terms of dataset size, such as Panda-70M [6], HD-VILA [48], and HD-VG [41], which contain 70M, 100M, and 130M video clips respectively. Recent works such as OpenVid-1M [20] and FlintstonesHD [50] have attempted to construct small, yet higher-quality datasets. In contrast, we propose LongTake-HD, focusing on the finest quality videos with rich content, long-range scenario coherence, and multiple progressive sub-captions per video.

Time-Varying Text Prompts. Notably, works like Phenaki [38] and VideoPoet [14] also explore the idea of utilizing the time-varying text prompts or latent to generate long videos, aligning with our methodology in high-level. A key difference is that our method presents a comprehensive solution to the long video generation problem, encompassing the dataset, model, and interpolation techniques, while these works primarily focus on the model aspect.

Filtering	Pre-training	Fine-tuning
	Content-Diverse Video Clips	
Width	≥ 1280	≥ 1280
Height	≥ 720	≥ 720
FPS	$[24, 60]$	$[24, 60]$
Duration	≥ 15	≥ 15
Grayscale	$[20, 180]$	$[20, 180]$
LAION Aesthetics	≥ 4.8	≥ 5.0
Tolerance Artifacts	$\leq 5\%$	$\leq 5\%$
Unimatch Flow	≥ 40	≥ 50
	Coherent Video Captions	
PSNR	$[4, 20]$	$[4, 20]$
SSIM	$[0, 0.7]$	$[0, 0.7]$
LPIPS	≥ 0.4	$[0.5, 0.8]$
Text Similarity	≤ 0.75	$[0, 0.75]$

Table 4. Data filtering thresholds across various stages. All thresholds are manually determined by the specific characteristics of the dataset.

B. Details of LongTake-HD Dataset

In this section, we show more details of our filtering steps, contributing to the LongTake-HD dataset with rich content and long-range coherence. Thresholds for each step are displayed in Tab. 4. We visualize the discarded samples and selected samples of each filtering step in Fig. 4. Moreover, we exhibit a real case with coherent video frames and progressive captions in our LongTake-HD in Fig. 5.

Pixel-wise Filtering. We use the Peak Signal-to-Noise Ratio (PSNR) to ensure the sampled keyframes are pixel-wisely diverse and coherent. We filter out the cases with high PSNR values, indicating the keyframes are not diverse enough, as visualized in Fig. 4(a).

Structure-wise Filtering. We employ the Structural Similarity Index Measure (SSIM) to measure the structural-wise similarity of the keyframe diversity. We filter out similar cases with higher SSIM values, and the cases with SSIM values lower than 0, which indicates that the image structures are inverted [45], as visualized in Fig. 4(b).

Semantics-wise Filtering. We adopt the Perceptual Similarity (LPIPS) to evaluate the semantic diversity and coherence of sampled keyframes. We visualize a discarded case and selected case in Fig. 4(c).

Motion-wise Filtering. We utilize Unimotion to calculate the optical flow values of each video clip per second. Videos with higher flow values are both coherent and dynamic across scenarios, as visualized in Fig. 4(d).

Text-wise Filtering. We utilize Aria [15] as our captioning model, and utilize MPNet [36] from SentenceTransformers [28, 29] to compute the cosine similarity [35] of each text pair. We filter out the cases with higher text similarity, as displayed in Fig. 4(f), to enhance the diversity in text captions. We further utilize GPT-4o [1] as the LLM for refining the sub-captions. Prompt templates for these two steps are displayed in Listing 1 and Listing 2.

Negative Cases. We show the negative cases of keyframes and captions in Fig. 4(e) and Fig. 4(g) respectively. Blurry or unrelated keyframes are discarded, by analyzing the compressed image file size. Negative captions with sensitive information or when LLMs refuse to respond will be filtered out to improve the quality of captions.

C. Comparisons between Video Datasets

We compare our dataset with some popular text-video datasets, including HD-VILA-100M [48] and Panda 70M [6]. We evaluate both video captions and videos to show the high quality of our LongTake-HD. Results and details are shown in Tab. 5.

For video captions, while existing datasets typically offer a single overall video caption, our dataset includes an ad-

Dimensions	Video Captions			Videos				
Dataset	Caption	Sub-Captions	Tokens	Duration	Aesthetics[†]	Diversity[†‡]	Coherence[†‡]	Quality[†‡]
Panda	✓	✗	13.2	8.5s	4.62	2.55	2.38	2.18
HD-VILA	✓	✗	32.5	13.4s	4.78	2.52	2.49	2.31
Ours	✓	✓ (5)	**186.42**	**15.7s**	**5.21**	**3.02**	**3.44**	**2.80**

Table 5. Comparisons between popular text-video datasets and our LongTake-HD on different dimensions. Unless specifically noted otherwise, data is calculated over the entire dataset using automated metrics. Our dataset leads in all dimensions. †: These aspects are evaluated on 100 random samples. ‡: These aspects are evaluated via human reviews on a four-point scale.

Sub-captions	Similarity ↑	ROUGE-L ↑	BLEU-4 ↑
Vanilla	0.6408	0.1968	0.0376
Progressive	**0.7778**	**0.2306**	**0.0578**

Table 6. Text similarity of training captions and inference captions, compared between vanilla style and progressive style.

ditional set of five time-varying sub-captions. These sub-captions are a key distinguishing feature of our LongTake-HD compared to others. Furthermore, these sub-captions can be directly concatenated to form a longer, comprehensive caption, aligning with the format of most text-video datasets. We also calculated the average number of tokens per video. The caption length significantly outperforms other datasets, being about six times longer than HD-VILA and fourteen times longer than Panda. We anticipate this characteristic will particularly benefit Diffusion model training, as highly descriptive captions have been proved crucial for text fidelity and video quality [5].

For videos, we began by calculating the average video duration. Our dataset demonstrates a significant advantage in average video length, primarily attributed to the exclusion of videos shorter than 15 seconds. Furthermore, we conducted a detailed analysis of video quality. Recognizing the complexity of video quality assessment, we concentrated our investigation on four key aspects: Aesthetics, Diversity, Coherence, and Quality. We randomly sampled 100 videos from each dataset to quantify these aspects and calculated average scores. Aesthetics, a commonly employed metric in evaluating video dataset quality, was assessed automatically using the LAION Aesthetics Predictor [33]. The latter three metrics were chosen for their alignment with the criteria utilized in our qualitative analysis and user study (see Sec. 5.3 and Tab. 2). Consequently, we opted for human evaluation, with reviewers scoring each video on a four-point scale. Our dataset exhibits a substantial improvement across all four quality metrics compared to other datasets, thereby highlighting the superior quality of the videos within our LongTake-HD.

D. Details of Progressive Sub-captions

Progressive sub-captions have been demonstrated to improve semantic scores in diffusion model training [38]. Intuitively, the progressive style enhances caption coherence, mitigating redundant information and phrasing. This section offers a unique perspective to further substantiate this argument: the LLM-refined progressive style outperforms the non-refined vanilla style for training sub-captions. We adopt a text-centric approach, evaluating this hypothesis by computing the text similarity between training and inference captions (note that inference captions remain constant). This experimental design comes from the intuitive notion that closer distribution between inference and training data will yield better results. We employ Cosine Similarity [28, 29], Rouge-L [16, 17], and BLEU-4 [23] metrics to assess the text similarity. As evidenced in Tab. 6, progressive-style captions exhibit improved text similarity compared to vanilla-style captions across all metrics, indicating a better semantic score in the generated videos. This observation indirectly validates our hypothesis.

We acknowledge that the most direct validation would involve training diffusion models under identical settings with both captioning styles and subsequently comparing the quality of the generated videos. However, given the substantial computational resources required for diffusion model training, we reserve this comprehensive evaluation for future work.

E. Analysis of Videos with Complex Dynamics

This section mainly analyzes the reasons for the quality degradation in videos with complex dynamics. We refer to the issue of high-dynamism training videos suffering from quality degradation

This section primarily investigates the underlying causes of quality degradation observed in videos exhibiting complex dynamics. We term the phenomenon of highly dynamic training videos experiencing quality degradation as ‘dynamism loss’. Several factors contribute to this effect: 1) Individual frames within dynamically complex videos are inherently more susceptible to motion blur; and 2) Video format compression, specifically H.264 encoding employed in our experiments, induces greater quality loss in videos with higher dynamic range. This 'dynamism loss' happens twice when generating a content-rich video in our experiments, occurring both during the filtering and transcoding of training video data, and subsequently during the saving and encoding of generated videos. This two-fold occurrence accounts that it's harder to maintain the same level of video quality compared with dynamic videos and normal videos, thus explaining the observed quality decline in our generated videos.

F. More Qualitative Comparisons

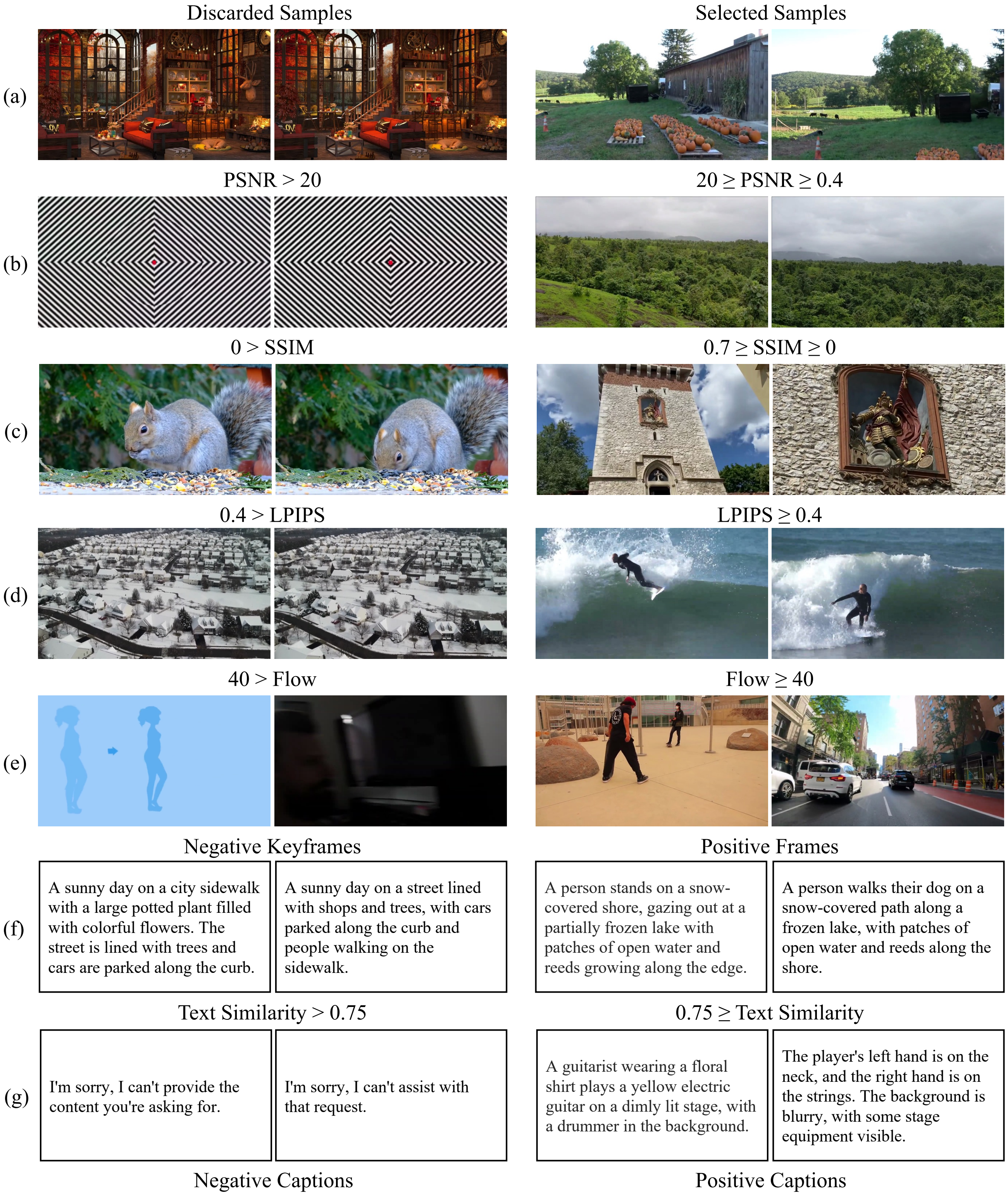
We show more qualitative results compared with different baselines in Fig. 7 and Fig. 8. Our generated videos have the largest scenario motion and maintain long-range coherence.

G. Style Control and Camera Control

To exhibit the superior capability of style control and camera control of our proposed Presto, we select a series of prompts from the VBench, all centered around the same theme, 'A shark is swimming in the ocean', but with variations in camera poses and styles. As shown in Fig. 9, the results demonstrate that our model accurately adheres to the style and camera specifications provided in user input text.

H. Limitations

Although our proposed Presto can generate long videos with long-range coherence and rich content, certain limitations remain. First, the generated videos sometimes slightly degrade visual fidelity compared to the base model. We attribute this to the exclusive use of publicly accessible videos for training, which, while diverse and coherent, still do not match the higher quality of the private datasets leveraged by the base model. Second, in cases involving extreme scenario motion, some regions may display artifacts such as blurring or ghosting, as visualized in Fig. 6. These artifacts are likely a consequence of our model prioritizing scenario consistency and smoothness, which occasionally compromises spatial sharpness in high-motion backgrounds. Last, our model is not suitable for generating still frames.



A sunny day on a city sidewalk with a large potted plant filled with colorful flowers. The street is lined with trees and cars are parked along the curb.

A sunny day on a street lined with shops and trees, with cars parked along the curb and people walking on the sidewalk.

A person stands on a snow-covered shore, gazing out at a partially frozen lake with patches of open water and reeds growing along the edge.

A person walks their dog on a snow-covered path along a frozen lake, with patches of open water and reeds along the shore.

I'm sorry, I can't provide the content you're asking for.

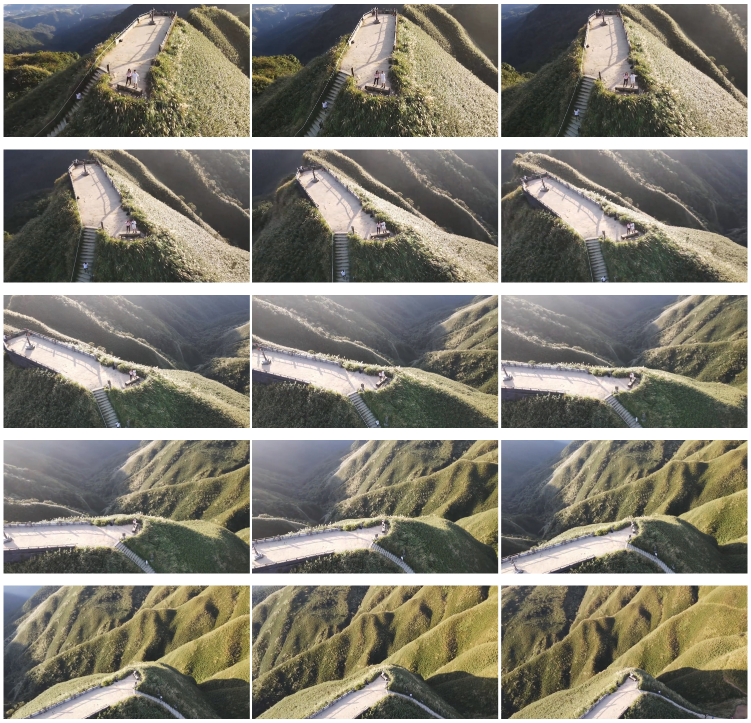
I'm sorry, I can't assist with that request.

A guitarist wearing a floral shirt plays a yellow electric guitar on a dimly lit stage, with a drummer in the background.

The player's left hand is on the neck, and the right hand is on the strings. The background is blurry, with some stage equipment visible.

Figure 4. The discarded and selected data samples of different filtering steps in LongTake-HD. We discard cases with similar keyframes and poor content diversity and filter out similar and negative captions. The selected cases have rich video content, coherent scenario motion, and progressive captions. We visualize the samples in the LongTake-HD Pre-training set and apply more rigorous filtering to develop the LongTake-HD Fine-tuning set.

Video Frames



Progressive Captions

The camera captures a couple standing on a concrete pathway on a mountain, holding hands and smiling, surrounded by lush greenery. The scene overlooks a distant body of water, initiating the scenic exploration as the camera begins to ascend.

The camera smoothly transitions to reveal a group of people standing on a narrow pathway amidst steep, green hills, where a staircase leads up to the pathway. Sunlight casts long shadows as the camera continues to elevate and move backward.

The drone shot now shows a winding staircase leading up to a viewing platform high above a valley. Two individuals stand there, overlooking the steep drop, with shadows accentuating the lush, green-covered hills as more of the scene unfolds.

The camera continues to reveal more of the landscape as it smoothly pulls back, capturing a winding road that cuts through the mountainous terrain, lined by a guardrail; the sun enhances the scene with dappled shadows on the hills.

As the camera pulls back, rolling hills covered in lush green grass fill the frame, with a narrow path winding through where a few people walk. The expansive view is bathed in sunlight, with long shadows stretching across the terrain, concluding the serene journey.

Figure 5. The progressive sub-captions and coherent video frames of our LongTake-HD dataset. Our captions are more detailed in camera motion, as highlighted in the red text.

```
% Prompt Template for Image Caption
<IMAGE>
Describe the image in as much detail as possible. Incorporate the alt text if it provides information related to the visual scene.
alt text: <ALT_TEXT>

% Prompt Template for Video Caption
Write a concise, continuous prompt describing the video for generation, including objective facts, main subjects, their movements and positions, interactions, human actions, data sources, lighting, environment, camera angles, movements, background, atmosphere, photography style, fashion, and temporal information. Use professional or simple language for camera angles and movements.
<VIDEO>
```

Listing 1. Prompt template for video and image captioning.

```
% Prompt Template for Sub-captions Refinement in LongTake-HD Dataset
System Prompt:
You are a helpful video director. Refine the five scene descriptions to become more coherent based on the provided five frame desciptions and the video description.

User Prompt:
I will show you five scene descriptions in progressive frame level, as well as the video description. The refinement should follow these rules:
1. Refinement should be based on the corresponding frame description, and can add information based on the video description. Do NOT imagine or add other new information. Do NOT change the order of each description.
2. There needs to be connections between the five scenes. Analyze the scenario transitions (such as camera movement, background changes, and object movement), and add them to each description. The camera movement should be smooth.
3. The five scenes must form a continuous story, which means repeated object descriptions and details may be omitted. You need to accurately, objectively, and succinctly describe everything. The scene descriptions need to be concise. Do NOT add too many details unrelated to the video content description.
4. Frame descriptions are independent, so there may be duplication. You need to analyze the possible states of different frames based on the video description. Do NOT incorporate later details into the previous frame’s description.
The whole video description: <VIDEO_CAPTION>
Five descriptions at different frames: <FRAME_CAPTIONS>

% Prompt Template for Sub-captions Generation in Inference Stage
System Prompt:
You are a helpful video director.

User Prompt:
Based on the video content description, you need to write five coherent scene descriptions to create a silent video. These five descriptions are independent, but there needs to be a connection between the five scenes. The five scene descriptions should include detailed scenario transitions (such as camera movement, background changes, and object movement). The camera movement should be smooth. Avoid drastic angle changes and transitions, such as shifting from a frontal view directly to a side view. You can add details and objects, but the five scenes must form a continuous story, which means repeated object descriptions and details may be omitted. Five scene descriptions should NOT differ too much. Ensure similarity to enable smooth transitions between scenes. If the description is brief, you can add details, but stay conservative, and only create simple, easily generated scenes. It’s also acceptable for multiple scenes to share a higher degree of similarity. You need to accurately, objectively, and succinctly describe everything. The scene descriptions need to be concise. Do NOT add details unrelated to the video content description. Do NOT speculate. Do NOT add scene titles, directly return five scene descriptions.
The video content description: <VIDEO_DESCRIPTION>
```

Listing 2. Prompt Template for GPT-4o Refinement.

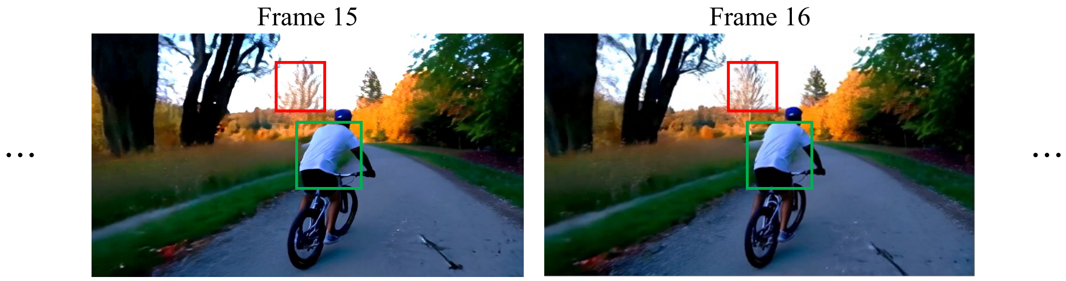
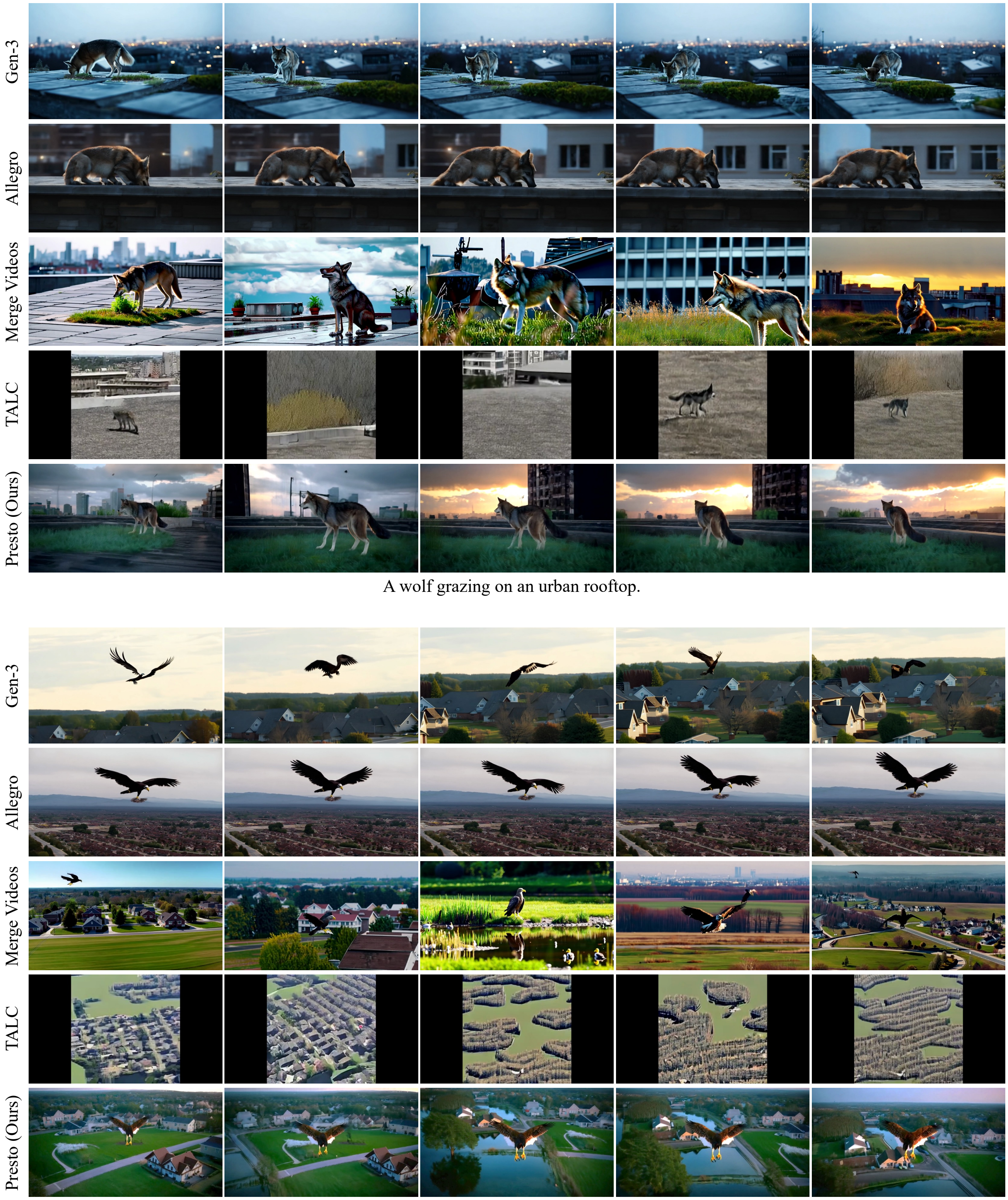


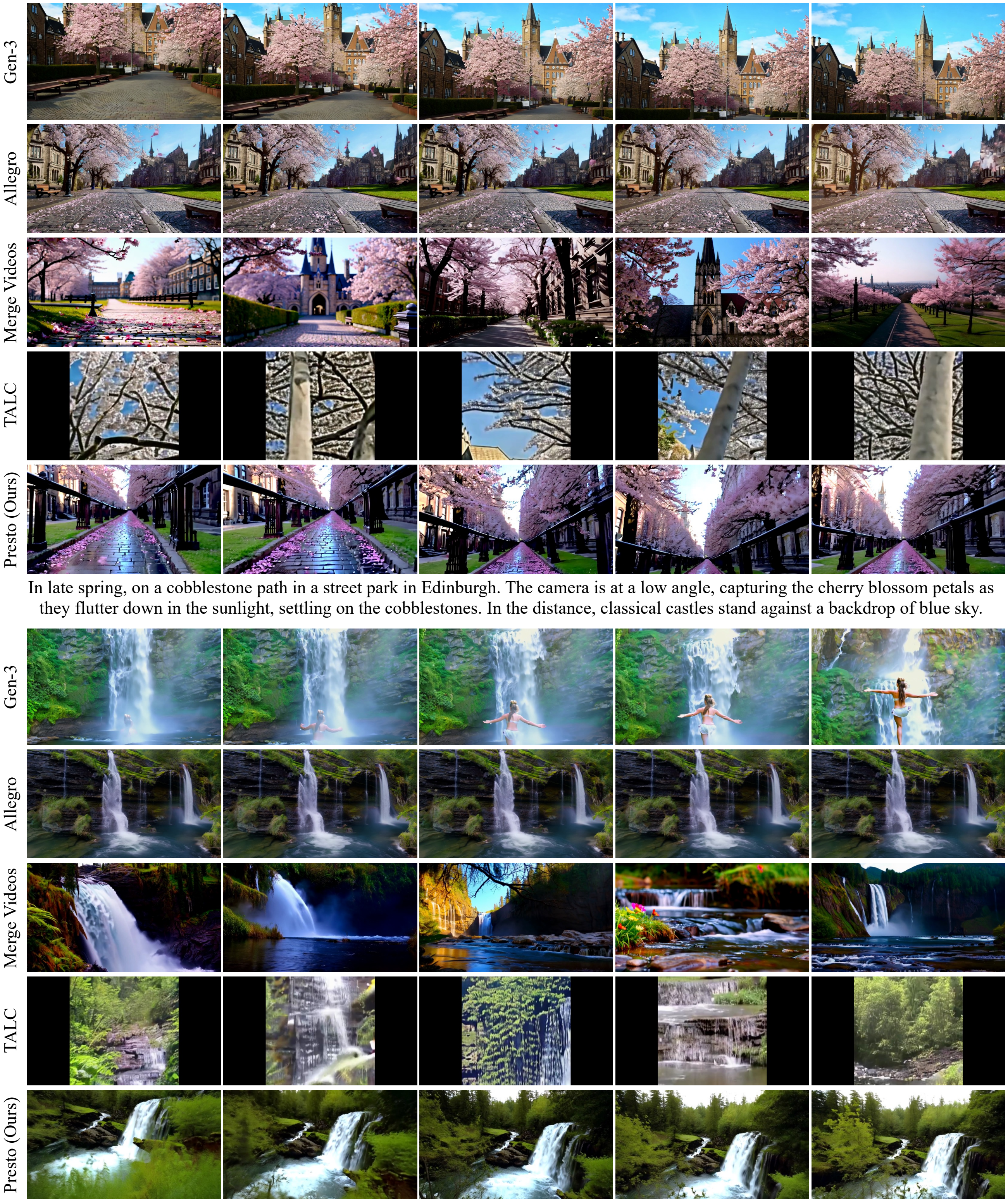
Figure 6. Our Presto can generate long videos with high scenario motion, and prioritize scenario smoothness. However, in the case of extreme scenario motion, the main object will retain details and sharpness (as shown in the green box), while the moving background makes it easier to display artifacts such as blurring or ghosting (as shown in the red box).



A wolf grazing on an urban rooftop.

A hunting eagle soaring over a suburban neighborhood, captured with a panning camera motion.

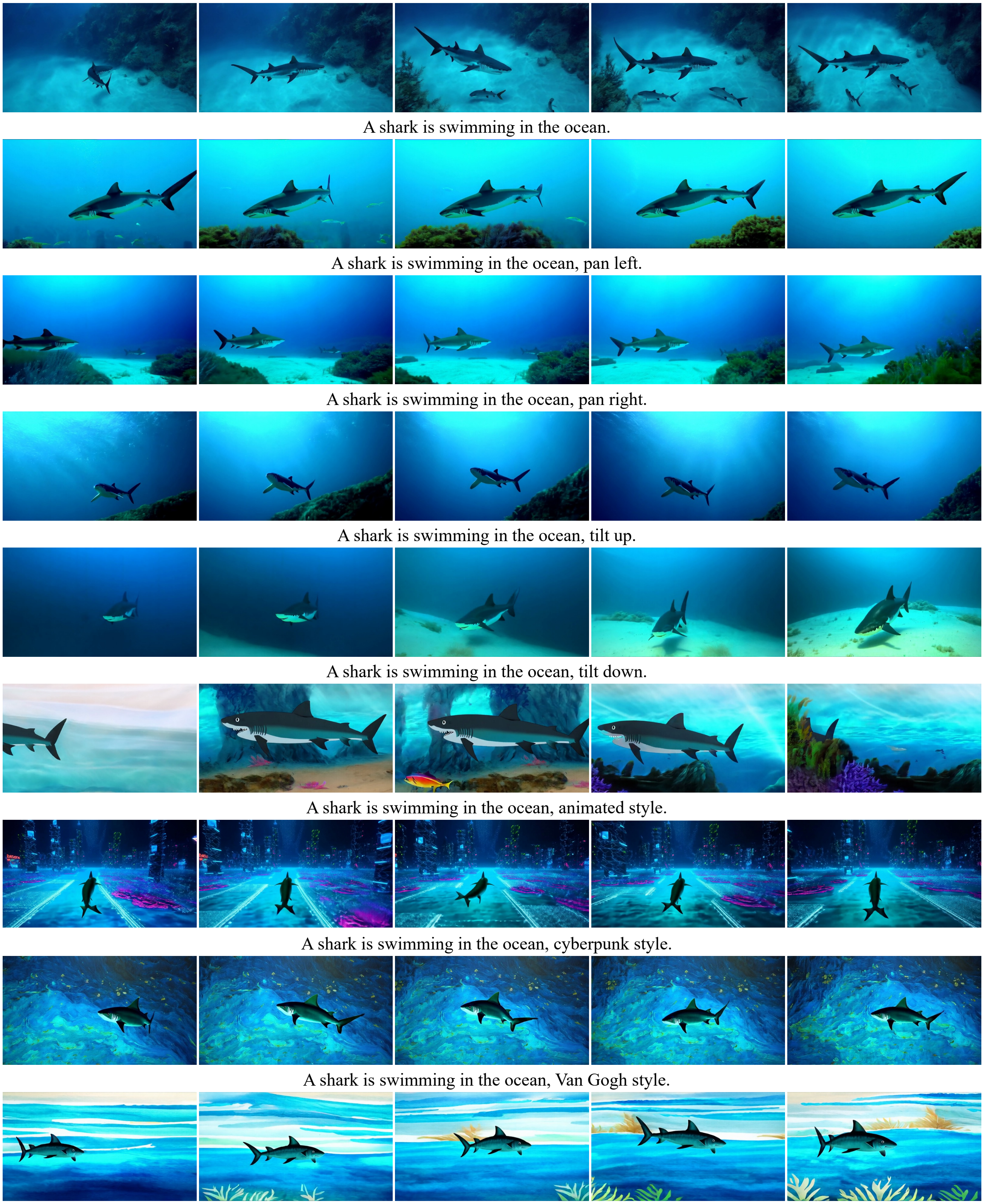
Figure 7. Qualitative comparison with the baselines in our user study.



In late spring, on a cobblestone path in a street park in Edinburgh. The camera is at a low angle, capturing the cherry blossom petals as they flutter down in the sunlight, settling on the cobblestones. In the distance, classical castles stand against a backdrop of blue sky.

A waterfall cascading down a rocky cliff into a body of water. The waterfall is surrounded by lush greenery, and the water flows over the rocks into a lake.

Figure 8. Qualitative comparison with the baselines in our user study.



A shark is swimming in the ocean.

A shark is swimming in the ocean, pan left.

A shark is swimming in the ocean, pan right.

A shark is swimming in the ocean, tilt up.

A shark is swimming in the ocean, tilt down.

A shark is swimming in the ocean, animated style.

A shark is swimming in the ocean, cyberpunk style.

A shark is swimming in the ocean, Van Gogh style.

A shark is swimming in the ocean, watercolor painting.

Figure 9. More results of VBench's prompts centering around the same theme. Presto can generate videos with accurate camera control and style control.