Appendix

This supplementary material provides additional implementation details, results to show the quality of our text-to-video generation method and its applications, and the importance of individual parts of our approach.

Sec. 7 provides additional implementation details about the frame generation and the motion field.

The quality of our text-to-video method without additional conditioning or specialization is investigated further in Sec. 8. To this end, qualitative results are presented and compared to the only publicly available state-of-the-art competitor CogVideo [15]. In order to analyze the relevance of our proposed procedures, several ablation studies are performed qualitatively and quantitatively.

Sec. 9 supplements our paper by elaborating on results for conditional text-to-video generation guided by pose information. In Sec. 10 we discuss more results of conditional text-to-video generation guided by edge information. Qualitative results and extensive ablation studies are presented.

Finally, Sec. 11 provides additional qualitative results and more comparison with a recent state-of-the art method Tune-A-Video [42] for the instruction-guided video editing task and compares to our Video Instruct-Pix2Pix method.

7. Additional Implementation Details

We utilize a classifier-free guidance [13] weight of 7.5 throughout all experiments. To obtain motion dynamics, we set $\delta_x = \delta_y = 1$. For text-to-video generation, we use $\lambda = 20$. In the conditional and specialized case and for Video Instruct-Pix2Pix, motion information is already provided by the conditioning. Therefore, we use $\lambda = 0$ in that case, so that each frame is generated using the same noise.

For the translation vector δ_k, we are using a symmetric padding and nearest neighbor interpolation. As the UNet with its convolutions and attentions (without positional encodings) is translation-equivariant, this allows to use our warping operation without sacrificing image quality.

8. Additional Experiments for Text-to-Video Unconditional Generation

8.1. Ablation Studies

We conduct additional ablation studies regarding background smoothing, cross-frame attention, latent motion and the number Δt of DDPM forward steps.

Timestep to apply motion on latents: Applying motion on the latent codes x_T (corresponding to $\Delta t = 0$) leads mainly to a global shift without any individual motion of the object, as can be seen in Fig. 10. It motivates our term *motion dynamics*, which influences the global camera scene and camera motion. The other extreme case, setting $\Delta t = 1000$, follows less the defined motion, so that the frame consistency is degraded, compare Fig. 15. The desired result is thus a trade-off between motion dynamics and local motions, which are modeled by local variations introduced by Δt steps of DDPM. In other words, Δt is a trade-off between global motion and local motion. We empirically set $\Delta t = 60$ in our method for all experiments, which provides good object motions (see Fig. 11).

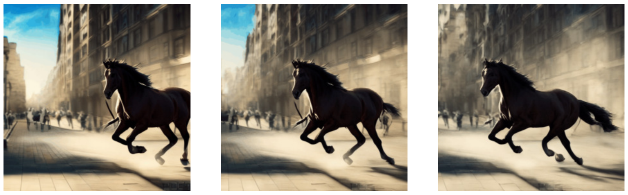


Figure 10: Generated frames using $\Delta t = 0$. As no DDPM is used, the frames are essentially translations of the first frame.

Background smoothing: We visualize the impact of using background smoothing in Fig. 12 and Fig. 13. When background smoothing is turned on, $\alpha = 0.6$ is used. When active, the background is more consistent and better preserved (see e.g. red sign in Fig. 13).

Cross-frame attention and motion latents: We present in an ablation study the importance of cross-frame attention and motion information on latent codes. The qualitative results are presented in Fig. 14. With the base model only, i.e. without our changes (first row), no temporal consistency is achieved. This is especially severe for unconstrained text-to-video generations. For example, the appearance and position of the horse changes very quickly, and the background is utterly inconsistent. Using our proposed motion dynamics (second row), the general concept of the video is preserved better throughout the sequence. For example, all frames show a close-up of a horse in motion. Likewise, the appearance of the woman and the background in the middle four figures (using ControlNet with edge guidance) is greatly improved.

Using our proposed cross frame attention (third row), we see across all generations improved preservation of the object identities and their appearances. Finally, by combining both concepts (last row), we achieve the best temporal coherence. For instance, we see the same background motifs and also about object identity preservation in the last four columns and at the same time a natural transition between the generated images.

We observe the same behaviour in the additional results on text-to-video presented in Fig. 16. Without cross-frame attention and without motion information on latents the scene differs from frame to frame, and the identity of the main object is not preserved. With motion on latents activated, the poses of the objects are better aligned. Yet, the appearance differs between the frames (e.g. when looking

	Vanilla	+MD	+MD+CF	+MD+CF+BS
textual faithfulness (↑)	**32.5**	31.64	31.51	31.49
frame consistency (↑)	81.7	84.3	90.8	**91.06**

Table 1: Ablation Study. We add motion dynamics (MD), cross-frame attention (CF) and background smoothing (BS) to the base model.

	Prev	Pref+First	First
textual faithfulness (↑)	31.26	**31.56**	31.49
frame consistency (↑)	87.7	90.3	**91.06**

Table 2: Ablation Study on cross-frame attention. We can attend on the previous frame (Prev), the previous and current frame (Prev+First) and on the first frame (First).

at the depicted dog sequence). The identity is much better preserved when cross-frame attention is activated. Also the background scene is more aligned. Finally, we obtain the best results when both, cross-frame attention and motion on latents are activated.

Finally, we analyse the impact of motion dynamics, cross frame attention and background smoothing quantitatively.

To this end, we evaluate the textual faithfulness and frame consistency as defined in Tune-A-Video [42]. The results presented in Tab. 1 show our proposed method significantly improves the frame consistency (+13%), which indicates the consistency in the global scene, and the preservation of the foreground object identity. The textual faithfulness is essentially kept (-0.3%).

Why not attending to the previous frame? Alternative to attending to the first frame in our cross-frame attention, it is possible to attend to the previous frame. However, we found that attending only to the previous frame leads to frame-wise error propagation, so that eventually the style of the video and identities are not kept. Quantitatively it degrades the frame consistency metric significantly (from 91.06 to 87.7). Similarly, using the first and the previous frames together slightly degrades the frame consistency (from 91.06 to 90.3) and is computationally more expensive. We obtain the best results (see also Tab. 2) by attending to the first frame.

8.2. Qualitative results

We provide additional qualitative results of our method in the setting of text-to-video unconditional synthesis. For high-quality generation, we append to each prompt presented in subsequent figures the suffix ”high quality, HD, 8K, trending on artstation, high focus”.

Fig. 17 shows qualitative results for different actions, e.g. ”skiing”, ”waving” or ”dancing”. Thanks to our proposed attention modification, generated frames are consistent across time regarding style and scene. We obtain plausible motions due to the proposed motion latent approach. As can be seen in Fig. 18, our method performs comparable or sometimes even better than a state-of-the-art approach CogVideo[15] which has been trained on a large-scale video data in contrast with our optimization-free approach. Fig. 18(a-b)&(e) show that generated videos by our method are more text-aligned than CogVideo, regarding the scene. Also the depicted motion is with higher quality in several video generation (e.g. Fig. 18(a)&(e)&(g)).

9. Text-to-Video with Edge Guidance

In Fig. 19 we present more video generation results by guiding our method with edge information. Results of specialized text-to-video, *i.e.* guided by edge information and using a specialized DreamBooth model are shown in Fig. 20. In Fig. 22 we show the effect of our cross-frame attention and motion in latents for text-to-video generation with edge guidance. As can be noticed when using CF-Attn layer the generation preserves the identity of the person better, and using motion in latents further improves the temporal consistency.

10. Text-to-Video with Pose Guidance

In Fig. 21 we present additional results of our method guided by pose information. In Fig. 23 we show the effect of our cross-frame attention and motion information in latents.

11. Video Instruct-Pix2Pix

In Fig. 24 we present additional results of instruct-guided video editing by using our approach combined with Instruct-Pix2Pix [2]. As shown in Figures 25 and 26 our method outperforms naive per-frame approach of Instruct-Pix2Pix and a recent state-of-the-art method Tune-A-Video [42]. Particularly, while being semantically aware of text-guided edits, Tune-A-Video has limitations in localized editing, and struggles to transfer the style and color information. On the other hand Instruct-Pix2Pix makes visually plausible edits on image level but has issues with temporal consistency. In contrast with the mentioned approaches our method preserves the temporal consistency when editing videos by given prompts.

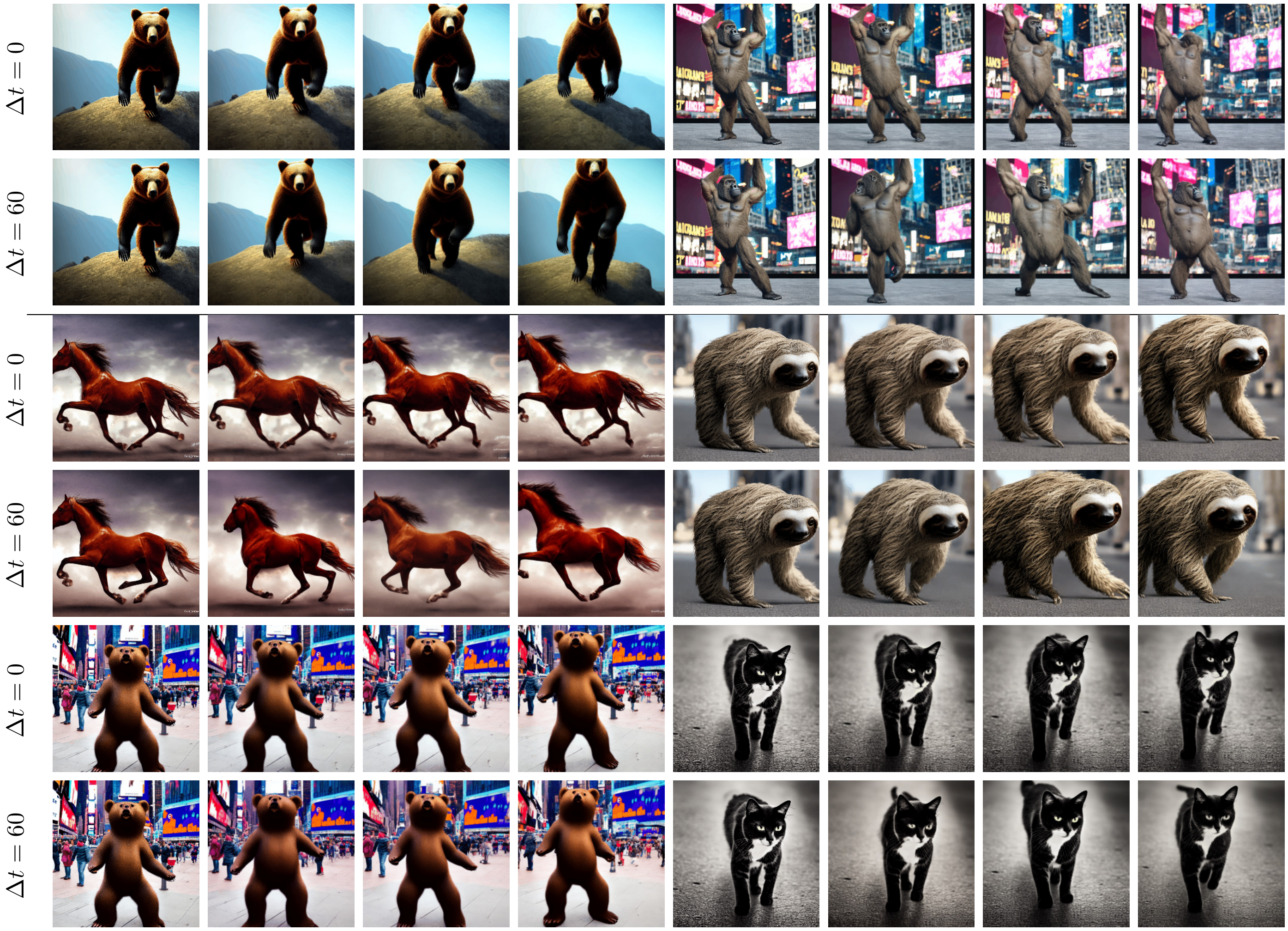
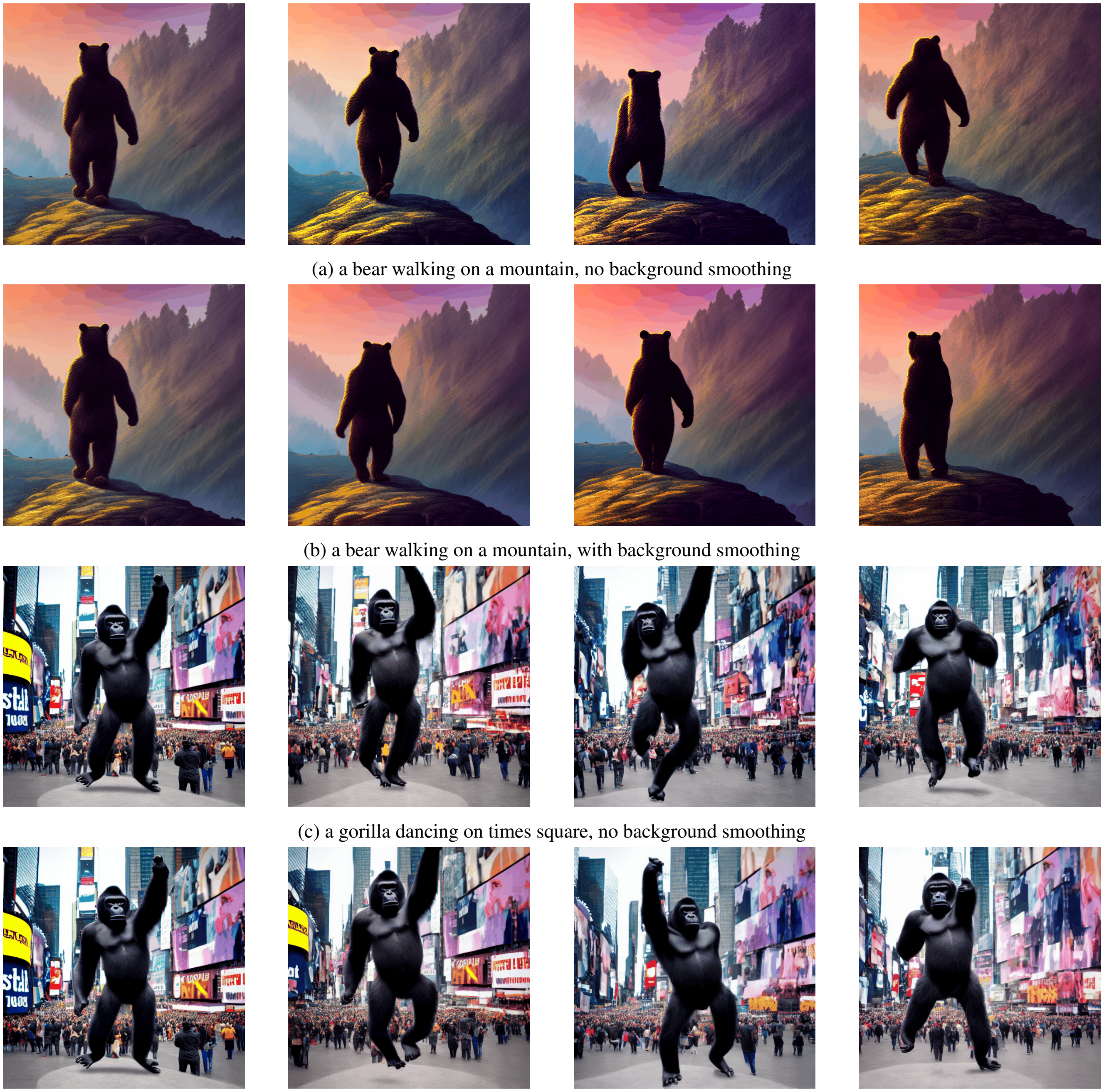


Figure 11: Ablation study on the number Δt of DDPM forward steps. The prompts to our method for text-to-video generation are (from left to right, top to bottom): “A bear walking on a mountain”, “a gorilla dancing on times square”, “A horse galloping on a street”, “a sloth walking down the street”, “A bear dancing on times square”, “A cat walking down the street.”



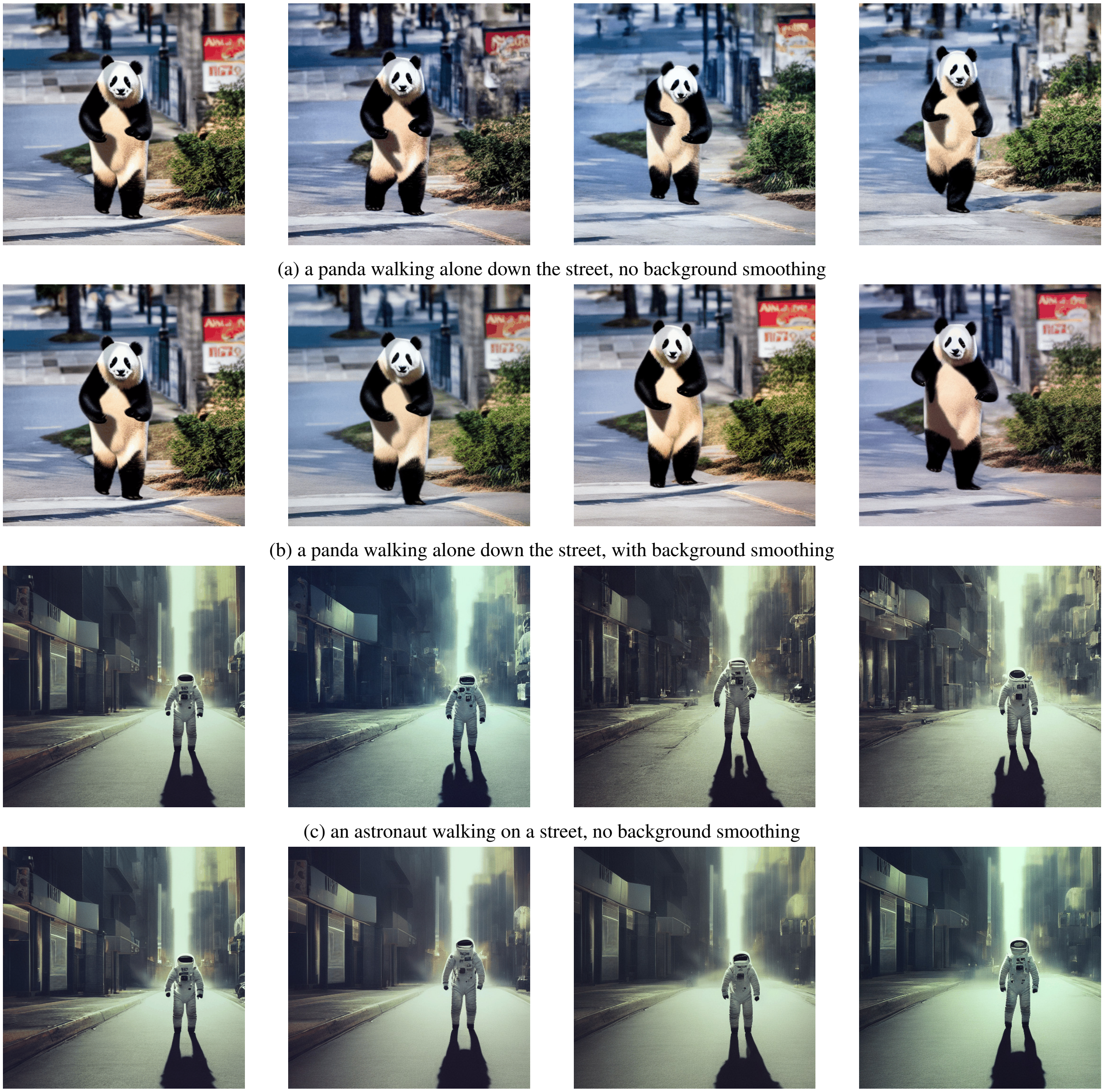
(a) a bear walking on a mountain, no background smoothing

(b) a bear walking on a mountain, with background smoothing

(c) a gorilla dancing on times square, no background smoothing

(d) a gorilla dancing on times square, with background smoothing

Figure 12: Ablation study on background smoothing.



(a) a panda walking alone down the street, no background smoothing

(b) a panda walking alone down the street, with background smoothing

(c) an astronaut walking on a street, no background smoothing

(d) an astronaut walking on a street, with background smoothing

Figure 13: Ablation study on background smoothing.

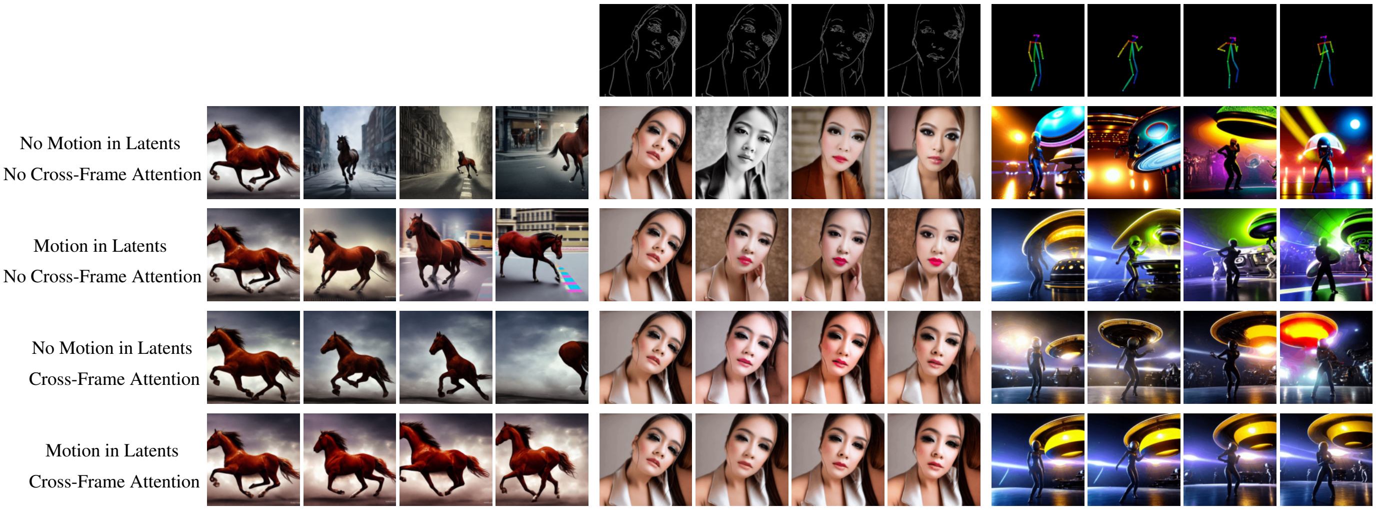


Figure 14: Ablation study showing the effect of our proposed components for text-to-video and text-guided video editing. Additional ablation study results are provided in the appendix.

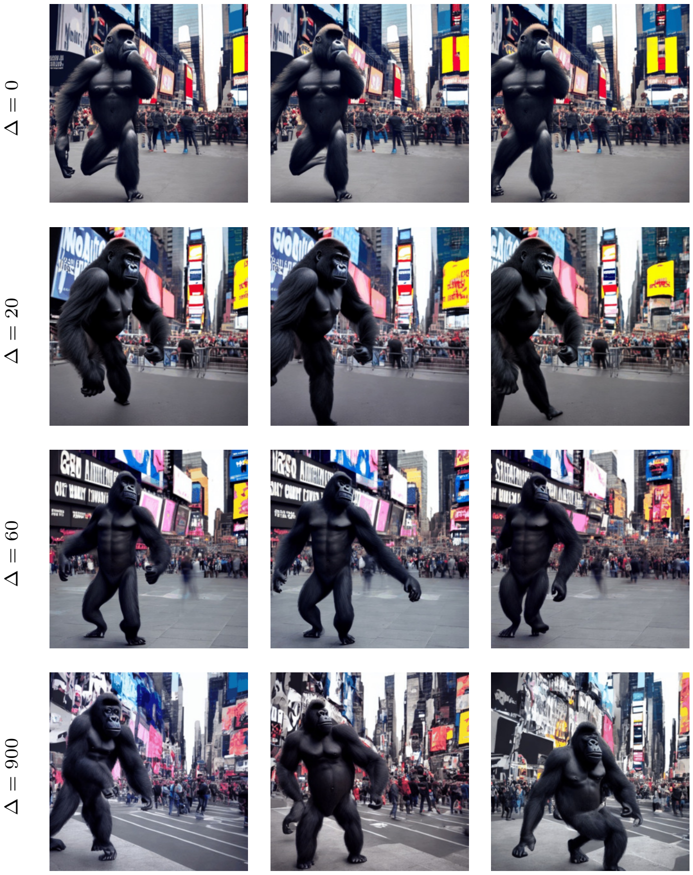


Figure 15: Ablation study on Δt. With larger Δt, we we can observe more object motion but less consistency in the background as we add more noise to the latent codes. The used prompt is: ”a gorilla dancing on times square”.

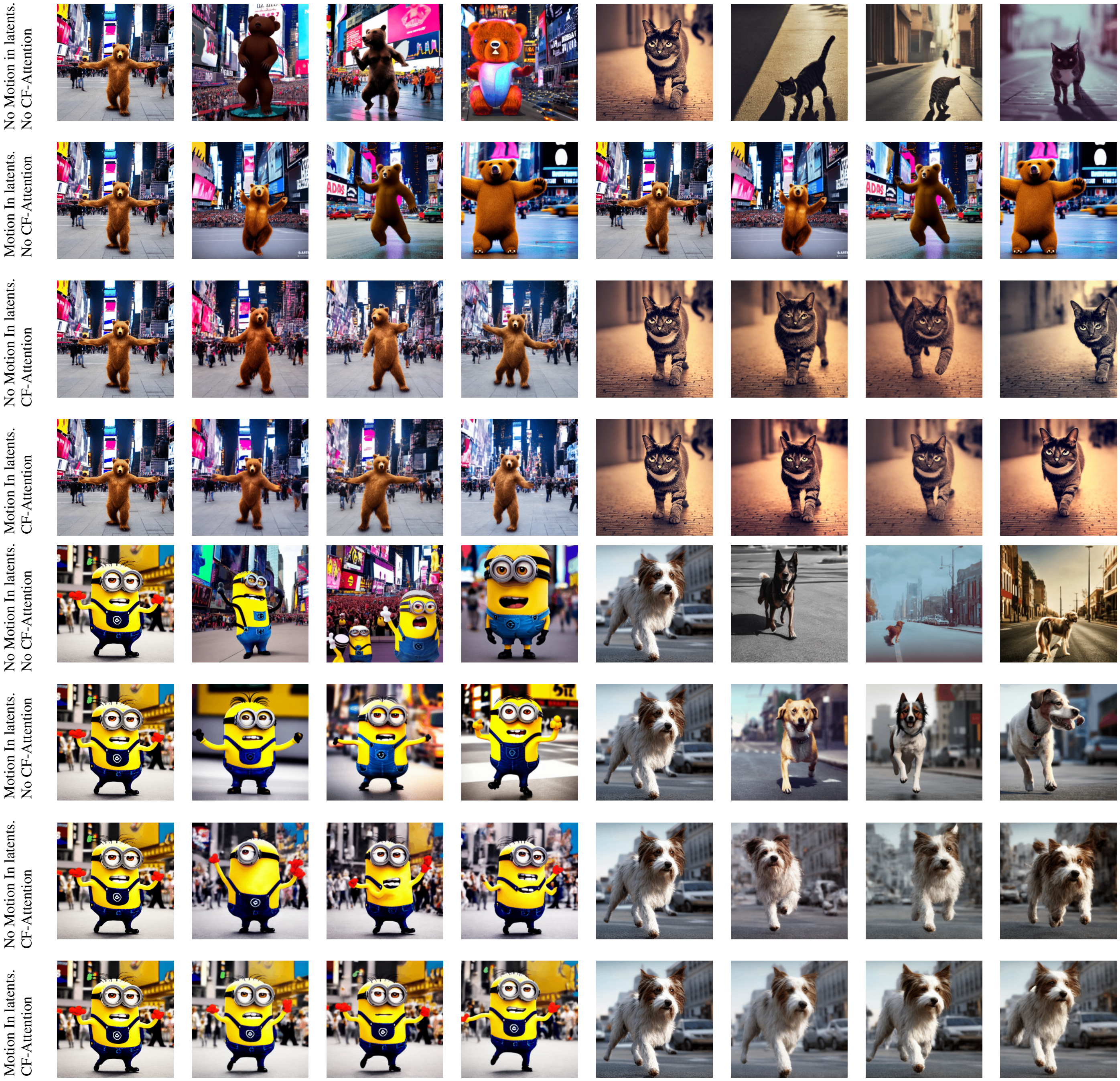
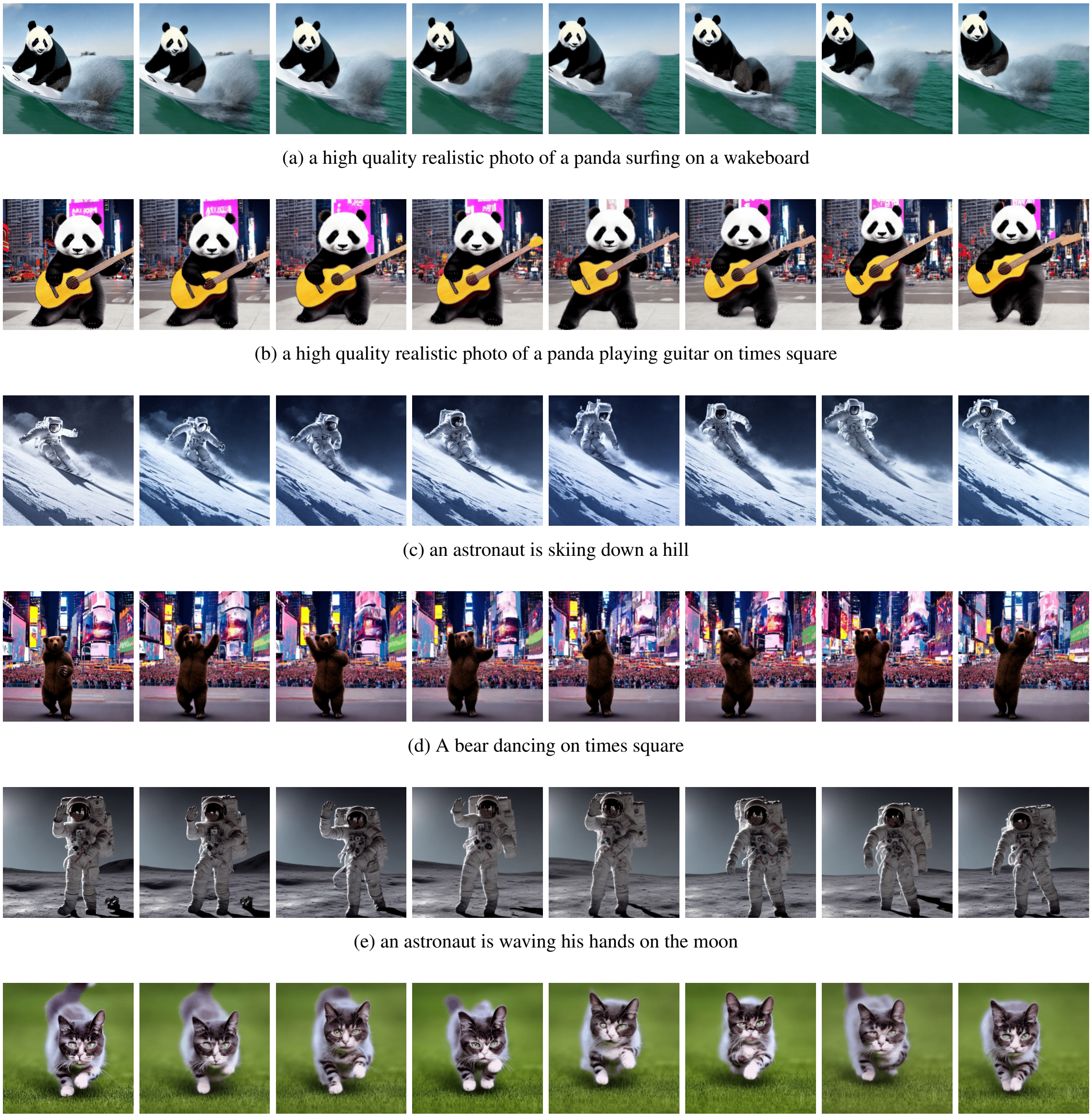

No Motion in latents. No CF-Attention
Motion In latents. No CF-Attention
No Motion In latents. CF-Attention
Motion In latents. CF-Attention
No Motion In latents. No CF-Attention
Motion In latents. No CF-Attention
No Motion In latents. CF-Attention
Motion In latents. CF-Attention

Figure 16: Ablation study on relevance of CF-Attention and Motion in latents. The left half of the rows 1-4 are generated with the prompt “A bear dancing on times square”. The right half of the rows 1-4 are generated with the prompt “A cat walking on a street”. The left half or the rows 5-8 are generated with the prompts “A Minion dancing on times square”. The right half of the rows 5 - 8 are generated with the prompt “A dog walking on a street.”



(a) a high quality realistic photo of a panda surfing on a wakeboard

(b) a high quality realistic photo of a panda playing guitar on times square

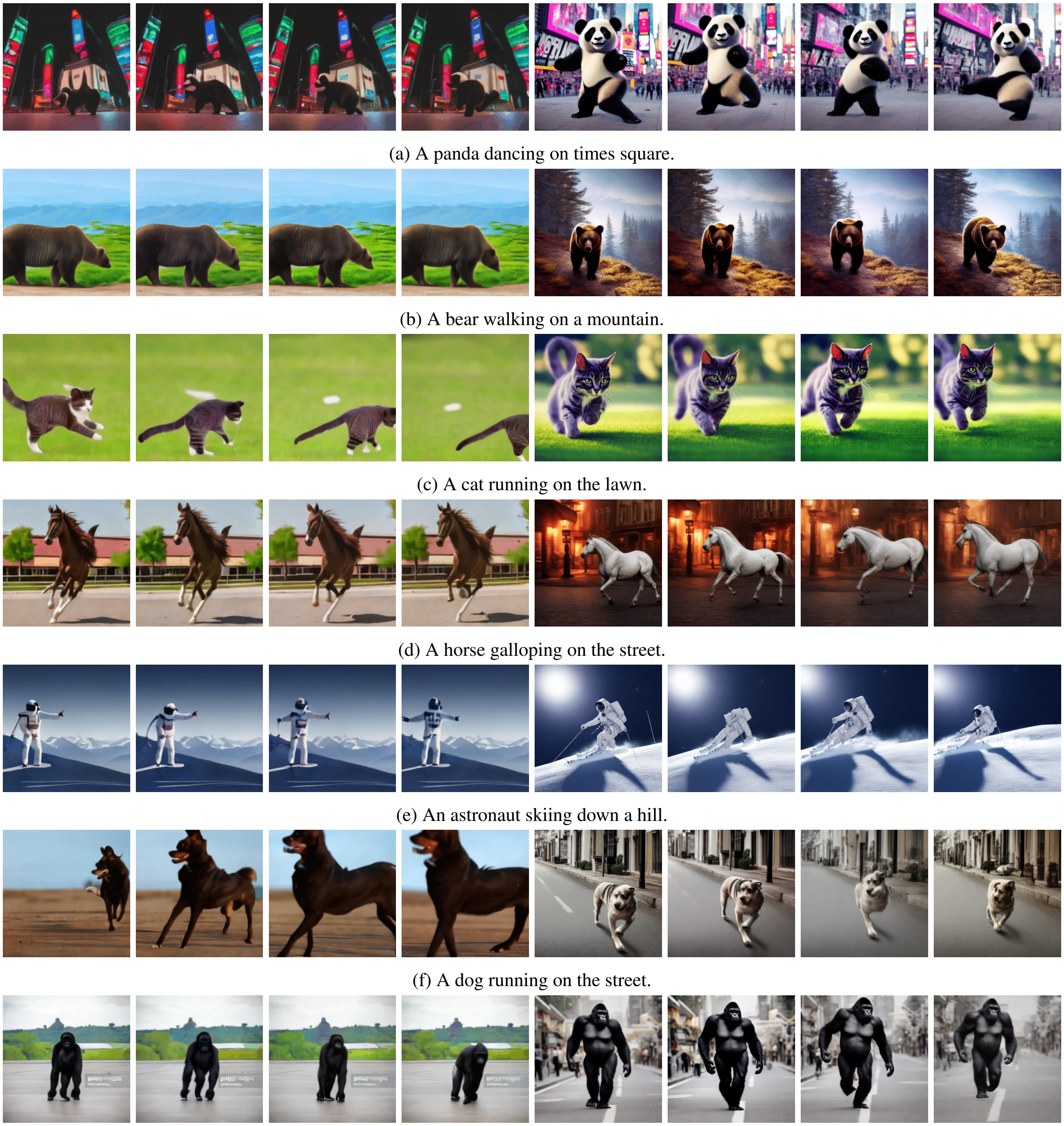
(c) an astronaut is skiing down a hill

(d) A bear dancing on times square

(e) an astronaut is waving his hands on the moon

(f) a high quality realistic photo of a cute cat running on lawn

Figure 17: Qualitative results of our text-to-video generation method for different prompts.



(a) A panda dancing on times square.

(b) A bear walking on a mountain.

(c) A cat running on the lawn.

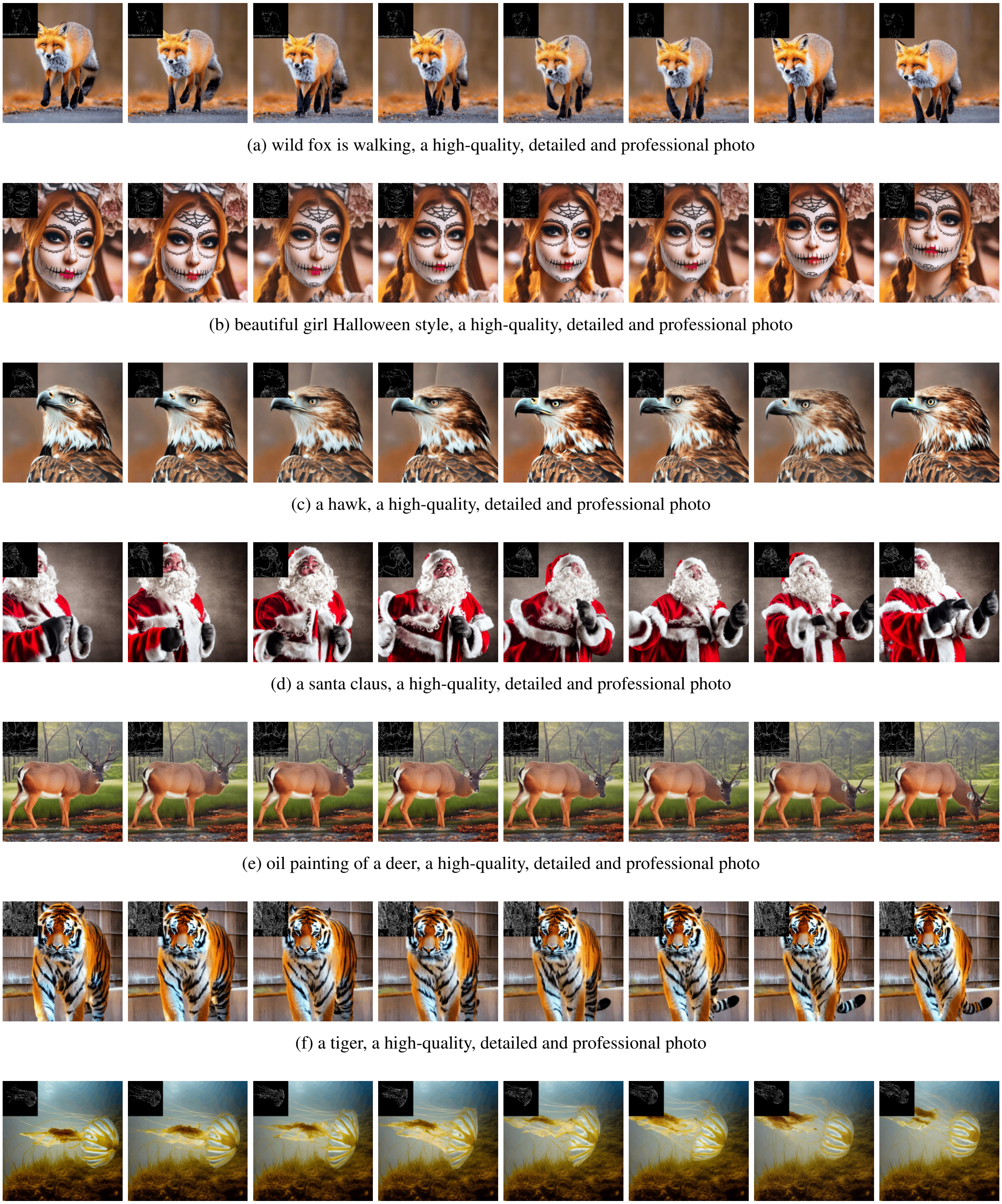
(d) A horse galloping on the street.

(e) An astronaut skiing down a hill.

(f) A dog running on the street.

(g) A gorilla walking down the street.

Figure 18: Qualitative comparison between CogVideo [15] (frames 1-4 in each row) and our method (frames 5-8 in each row).



(a) wild fox is walking, a high-quality, detailed and professional photo

(b) beautiful girl Halloween style, a high-quality, detailed and professional photo

(c) a hawk, a high-quality, detailed and professional photo

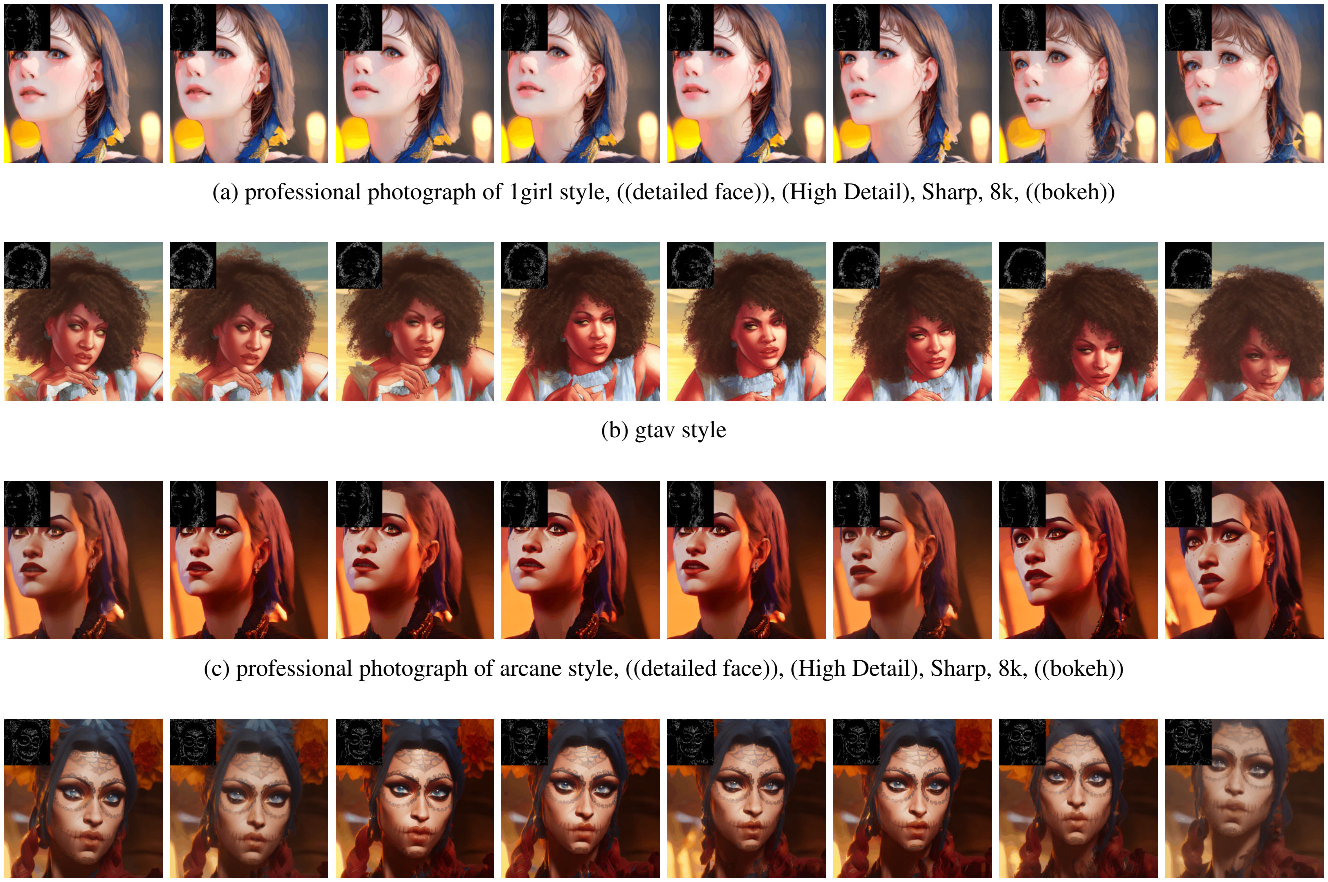
(d) a santa claus, a high-quality, detailed and professional photo

(e) oil painting of a deer, a high-quality, detailed and professional photo

(f) a tiger, a high-quality, detailed and professional photo

(g) a jellyfish, a high-quality, detailed and professional photo

Figure 19: Conditional generation of our method with edge control.



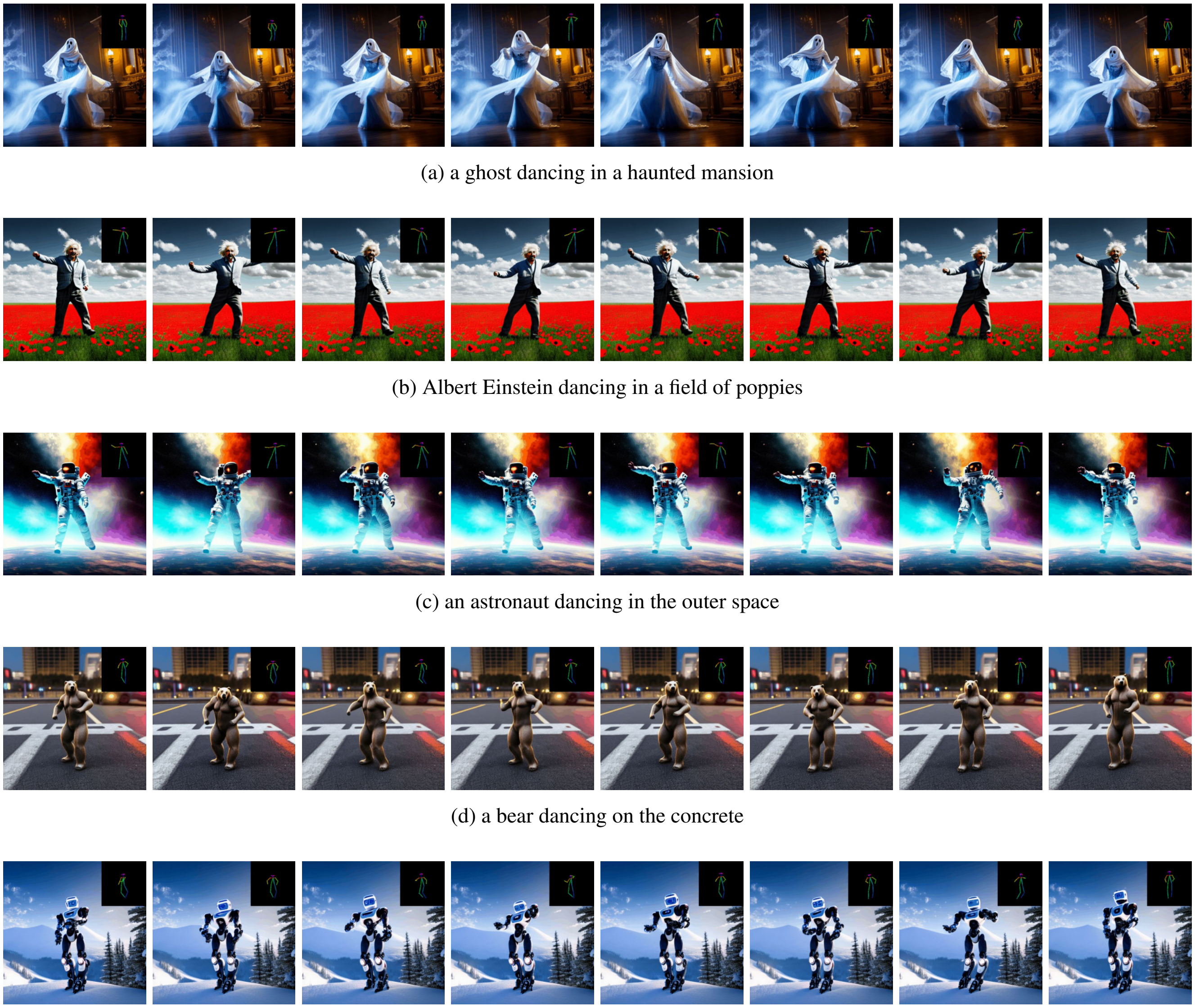
(a) professional photograph of 1girl style, ((detailed face)), (High Detail), Sharp, 8k, ((bokeh))

(b) gtav style

(c) professional photograph of arcane style, ((detailed face)), (High Detail), Sharp, 8k, ((bokeh))

(d) A stunning intricate full color portrait of arcane style, epic character composition, sharp focus, natural lighting, subsurface scattering, f2, 35mm, film grain

Figure 20: Conditional generation with edge control and DreamBooth [31] models. The keywords "1girl style", "gtav", "arcane style", and "arcane style" correspond to the personalized models of Anime Style, GTA style, and Arcane Style.



(a) a ghost dancing in a haunted mansion

(b) Albert Einstein dancing in a field of poppies

(c) an astronaut dancing in the outer space

(d) a bear dancing on the concrete

(e) a robot dancing on top of a snowy mountain

Figure 21: Conditional generation with pose guidance.

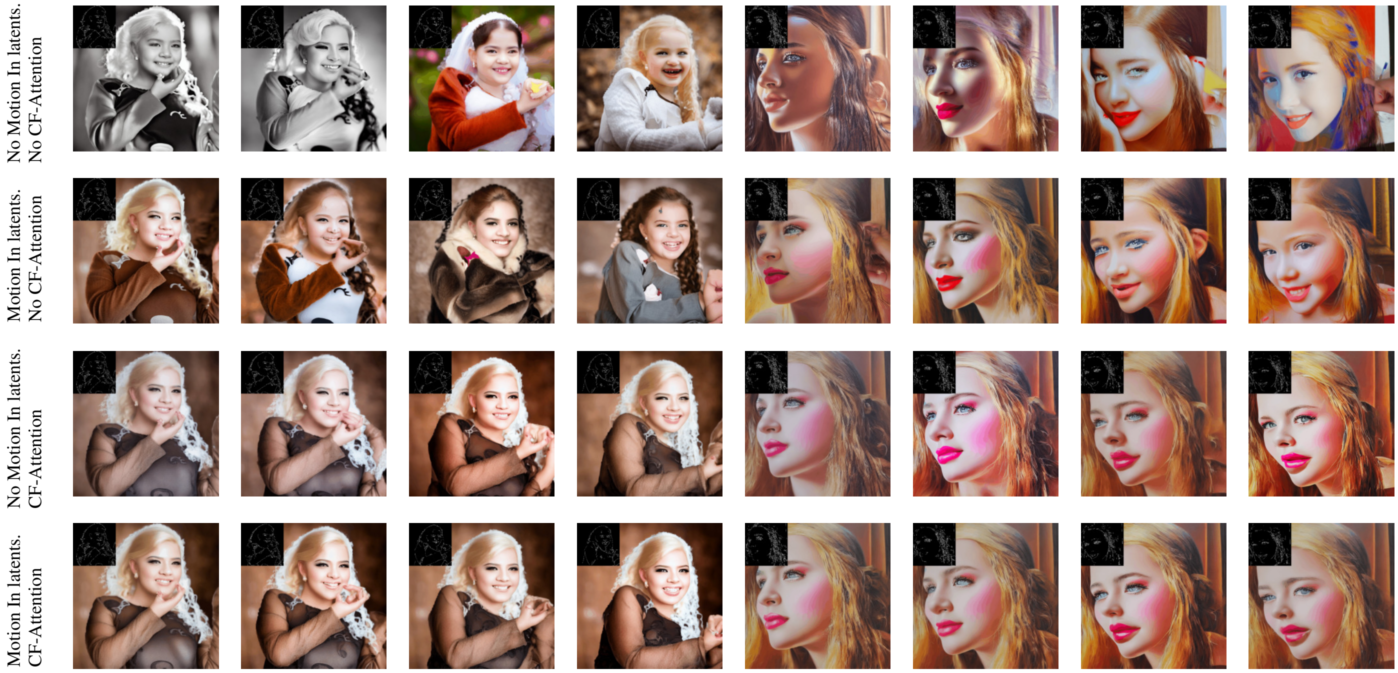


Figure 22: Ablation study on our CF-Attn block and adding motion information in latents for edge-guided video generation. The left half of each row is generated with the text prompt “A beautiful girl”. The right half of each row is generated with the text prompt “oil painting of a girl.”

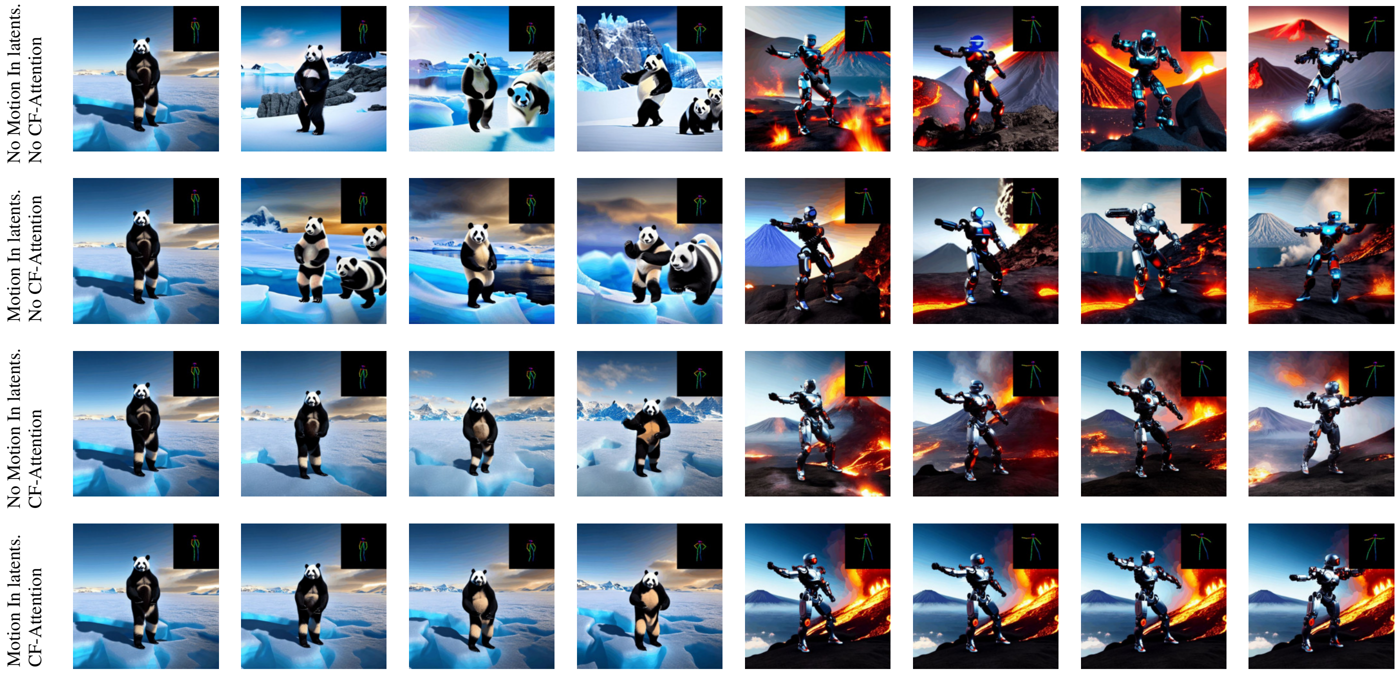
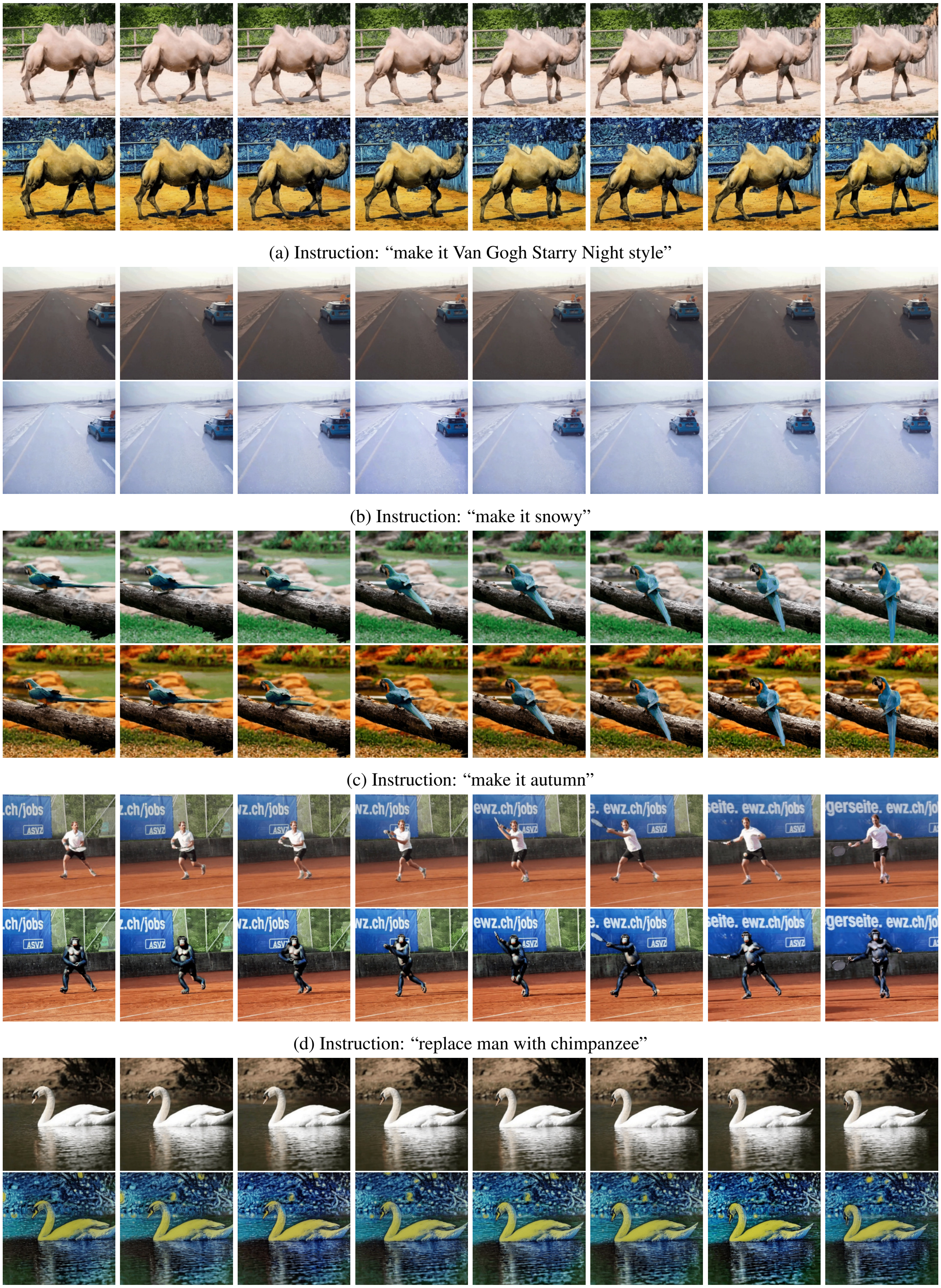


Figure 23: Ablation study on our CF-Attn block and adding motion information in latents for pose-guided video generation. The left half of each row is generated with the text prompt “a panda dancing in Antarctica”. The right half of each row is generated with the text prompt “a cyborg dancing near a volcano”.



(a) Instruction: “make it Van Gogh Starry Night style”

(b) Instruction: “make it snowy”

(c) Instruction: “make it autumn”

(d) Instruction: “replace man with chimpanzee”

(e) Instruction: “make it Van Gogh Starry Night style”

Figure 24: Text-guided video editing using our model combined with Instruct-Pix2Pix [2]. Instructions for editing are described under each video sequence.

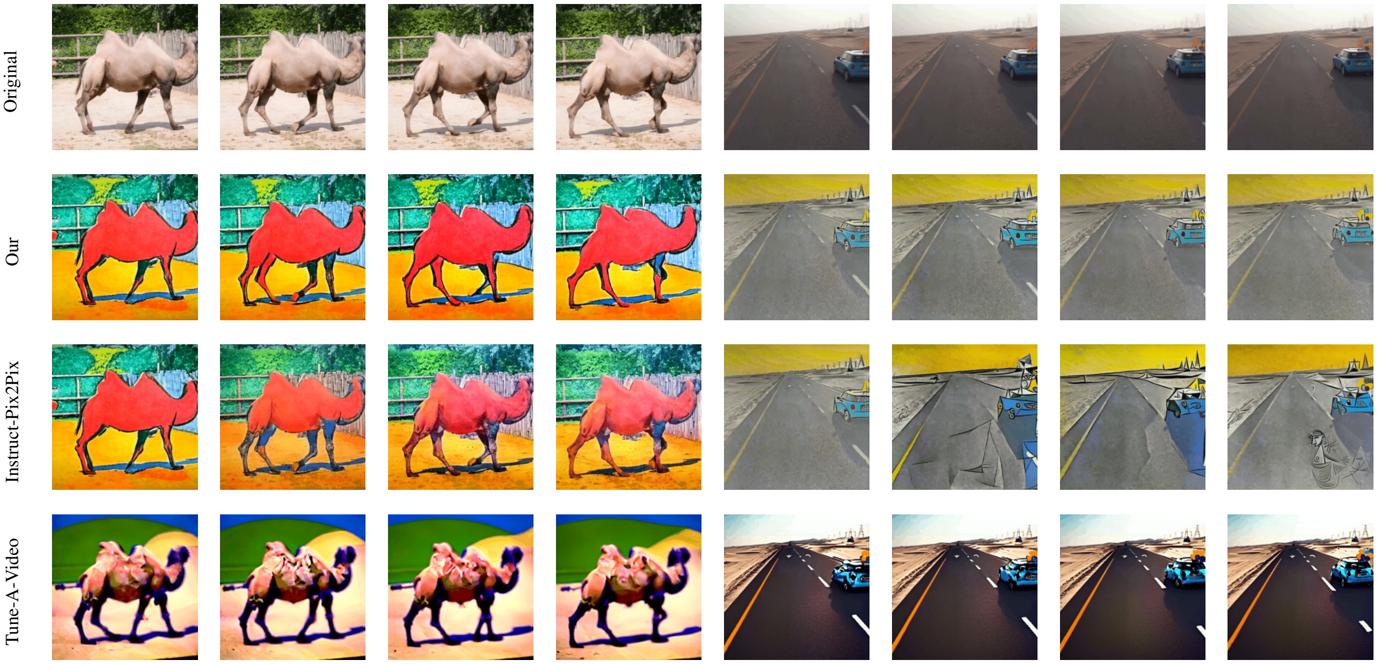


Figure 25: Text guided video editing using our method compared to Instruct-Pix2Pix [2] frame-by-frame and Tune-A-Video [42]. The left half of each row is generated with the instruction "make it Expressionism style" for our method and Instruct-Pix2Pix and "a camel walking, **in Expressionism style**" for Tune-A-Video. The right half of each row is generated with the text prompt "make it Picasso style" and "a mini cooper riding on a road, **in Picasso style**" respectively.

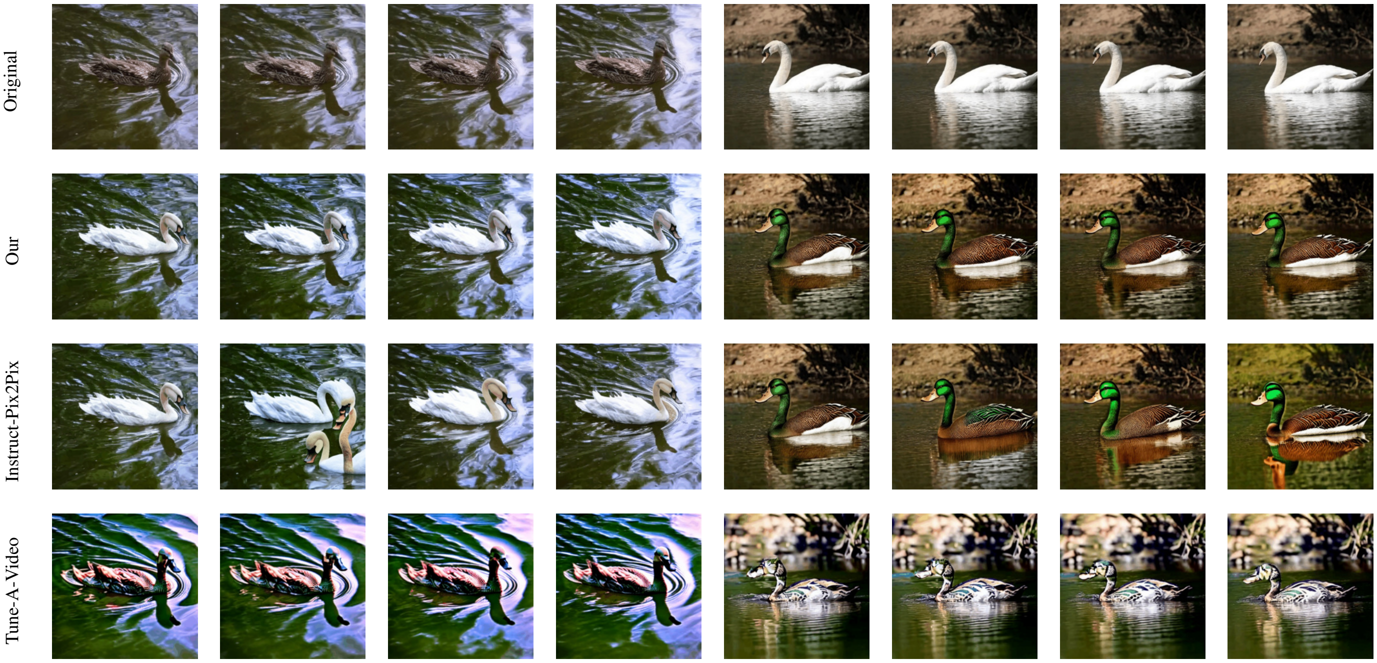


Figure 26: Text guided video editing using our method compared to Instruct-Pix2Pix [2] frame-by-frame and Tune-A-Video [42]. The left half of each row is generated with the instruction "replace mallard with swan" for our method and Instruct-Pix2Pix and "a **swan** swimming on water" for Tune-A-Video. The right half of each row is generated with the text prompt "replace swan with mallard" and "a **mallard** swimming on water" respectively.