CAMEL: CAusal Motion Enhancement tailored for Lifting Text-driven Video Editing

Guiwei Zhang[1]*, Tianyu Zhang[2]*, Guanglin Niu[3]†, Zichang Tan[4], Yalong Bai[2], Qing Yang[2]
[1] School of Computer Science and Engineering, Beihang University. [2] Du Xiaoman Financial.
[3] Institute of Artificial Intelligence, Beihang University.
[4] Department of Computer Vision Technology (VIS), Baidu Inc.
{zhangguiwei,beihangngl}@buaa.edu.cn, tianyu1949@gmail.com
tanzichang@baidu.com, ylbai@outlook.com, yangqing@duxiaoman.com

1. Additional Visualization Results on Multi-object Editing

Fig. 1 presents the superior performance of CAMEL in multi-object editing. In comparison to Tune-A-Video, which is plagued by serious flickering issues, CAMEL demonstrates significantly improved performance in enhancing both visual consistency and motion coherence for each object within a scene. Unlike Tune-A-Video, which utilizes global-scale temporal self-attention to model motion patterns and appearance content in an intricately intertwined fashion, CAMEL operates within a more targeted scope. Specifically, CAMEL focuses on enhancing the motion coherence of disentangled high-frequency components within each filtering window. Simultaneously, it ensures the preservation and generalization of low-frequency components that represent static appearance content to diverse creative textual prompts. This constrained approach proves beneficial in locally improving motion coherence and content consistency across overlapping filter windows. The advantages of this paradigm become particularly evident in scenarios involving multi-object editing, where it markedly outperforms the state-of-the-art approaches.

2. Additional Qualitative Results

We further perform a detailed visual comparison of CAMEL with the state-of-the-art approaches. In Fig. 2 (c), while Tune-A-Video effectively replaces the subject in the video template with “a furry rabbit” in accordance with the textual prompt, it falls short in transferring the essential motion pattern of “dives into a pool” from the original video template to the generated video. Besides, due to the suboptimal efficacy of canny conditions, ControlVideo exhibits limitations in maintaining content generalization and effectively capturing motion patterns. In contrast, our CAMEL framework demonstrates a remarkable capability to accurately capture motion patterns from the video template and effectively transfer them to a different subject “a furry rabbit”. In Fig. 2 (b), we observe that Tune-A-Video struggles to maintain the motion coherence inherent in the original video template. Additionally, it falls short in generating a video featuring a new subject, specifically “a rainbow colored squirrel.” A plausible explanation for these shortcomings is the underlying operational paradigm of Tune-A-Video, which involves learning motion and appearance in a deeply intertwined manner. This complex interplay often results in the network either overfitting to appearance content — thereby failing to accurately capture the essential motion patterns — or concentrating solely on motion patterns, which compromises the generation of content aligned with creative textual prompts. In contrast, our CAMEL is engineered to enhance the motion coherence of high-frequency components that capture contextualized motion patterns, while preserving the generalization of low-frequency components that represent appearance content. The above results highlight the comprehensive proficiency of CAMEL in balancing the intricacies of motion coherence and appearance generalization, showcasing its advanced performance in video editing.

3. Additional Quantitative Results

We further compare our CAMEL with the state-of-the-art video editing approaches: TokenFlow [1] and FateZero [2]. In Tab. 1, CAMEL exhibits superior performance against TokenFlow in terms of video-text alignment and frame consistency on the TGVE-DAVIS dataset. In Tab. 2, **to validate the efficacy of CAMEL in enhancing long-term motion coherence, we sample 32 uniform frames from the input video, in contrast to the 8-frame sampling used in FateZero**. The results show that CAMEL consistently

*Equal contribution.
†Corresponding author

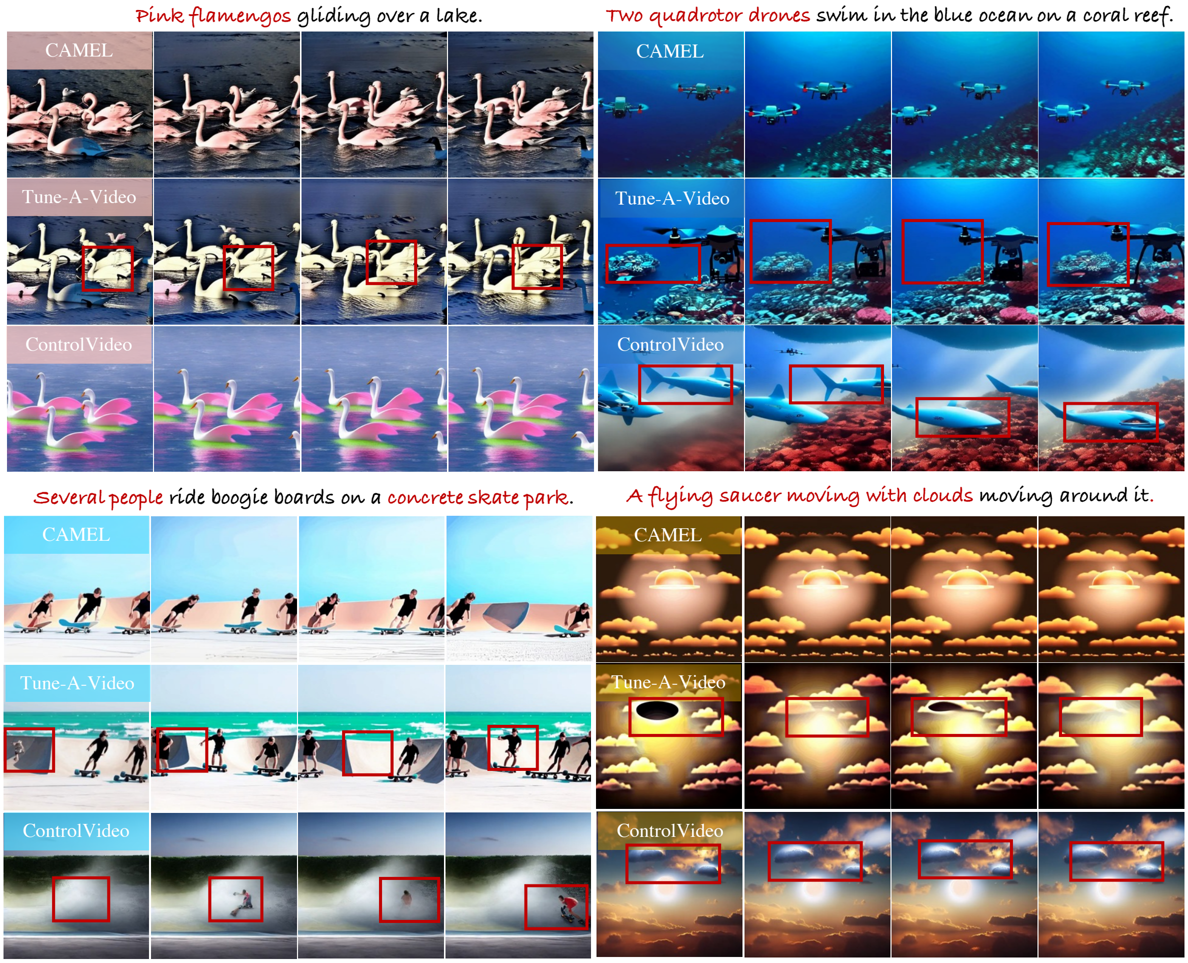
Pink flamengos gliding over a lake.
Two quadrotor drones swim in the blue ocean on a coral reef.
CAMEL
Tune-A-Video
ControlVideo
CAMEL
Tune-A-Video
ControlVideo
Several people ride boogie boards on a concrete skate park.
A flying saucer moving with clouds moving around it.
CAMEL
Tune-A-Video
ControlVideo
CAMEL
Tune-A-Video
ControlVideo

Figure 1. Illustration of our method's superior performance in multi-object editing.

Method	Video-Text Alignment Object	 Background	 Style	 Multiple
TokenFlow	35.01/24.08	35.35/26.10	33.35/26.49	32.89/25.13
AnimateDiff	35.18/25.71	/23.80	33.57/25.85	32.96/24.66
CAMEL	**36.02/28.41**	**35.97/27.85**	**34.97/28.66**	**34.68/28.35**

Method	Frame Consistency Object	 Background	 Style	 Multiple
TokenFlow	92.88	92.75	93.34	93.18
AnimateDiff	88.79	90.08	91.10	89.68
CAMEL	**93.94**	**93.27**	**93.72**	**93.40**

Table 1. Quantitative Results in Frame Consistency and Video-Text Alignment **(UMTScore/CLIPScore)** on the TGVE-DAVIS.

Method	Video-Text Alignment UMTScore	 CLIPScore	Frame Consistency
FateZero	34.91	24.46	92.77
CAMEL	**36.02**	**28.41**	**93.94**

Table 2. Comparisons with FateZero on the local attribute editing.

outperforms FateZero in terms of local attribute editing on the TGVE-DAVIS dataset. Unlike FateZero, which improves temporal consistency by simply warping the middle frame, we introduce a novel CAM-Attn mechanism, effectively paired with a causal motion filter. This enables a more continuous and accurate representation of motion patterns, particularly over extended time periods.

References

[1] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1

[2] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 1

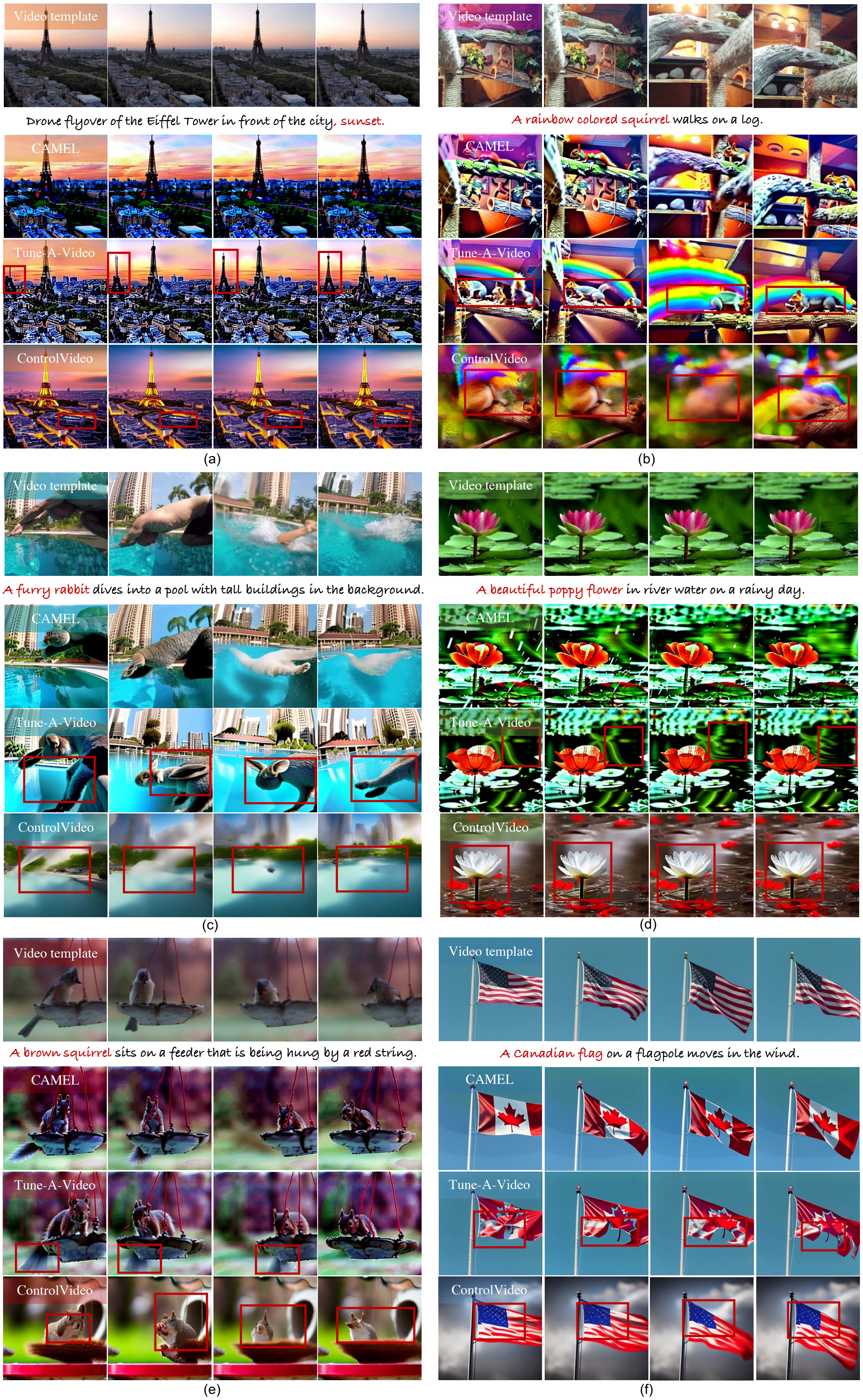
Video template
Drone flyover of the Eiffel Tower in front of the city, sunset.
CAMEL
Tune-A-Video
ControlVideo
(a)
Video template
A rainbow colored squirrel walks on a log.
CAMEL
Tune-A-Video
ControlVideo
(b)
Video template
A furry rabbit dives into a pool with tall buildings in the background.
CAMEL
Tune-A-Video
ControlVideo
(c)
Video template
A beautiful poppy flower in river water on a rainy day.
CAMEL
Tune-A-Video
ControlVideo
(d)
Video template
A brown squirrel sits on a feeder that is being hung by a red string.
CAMEL
Tune-A-Video
ControlVideo
(e)
Video template
A Canadian flag on a flagpole moves in the wind.
CAMEL
Tune-A-Video
ControlVideo
(f)

Figure 2. Qualitative comparisons of our CAMEL framework against the state-of-the-art approaches.