

Supplementary Materials: MagicVFX: Visual Effects Synthesis in Just Minutes

Anonymous Authors

1 INTRODUCTION

In this document, we provide additional materials to support our main submission. We first present a detailed introduction to the dataset we construct. Subsequently, we analyze potential research directions that may warrant attention in future AIGC-based visual effects synthesis tasks. Finally, we give more results of our experiments.

2 DATASET

Dataset collection plays a crucial role in assessing model performance. Unlike supervised tasks such as visual reconstruction and visual retrieval, where evaluation datasets can be readily obtained, the outcomes of visual effects synthesis are often subject to human subjective judgment. Consequently, we focus solely on gathering data that can serve as input, without taking into account the ground truth for the synthesized results.

Our VFX-307 dataset comprises two distinct types of data: base videos and effect videos. The base videos serve as a common input for both SRE and SNRE synthesis paradigms, while the special effect videos are exclusively utilized as inputs for the SRE paradigm. Further, we construct sample sets suitable for SRE and SNRE based on these two types of data, respectively. The detailed procedures for processing and constructing the dataset have been thoroughly described in the main text. Therefore, in this section, we concentrate on analyzing the content of our dataset.

Effect Videos. We gather a collection of visual effects videos created and uploaded by professionals through online platforms, categorizing them into 5 categories based on their content and application, with detailed classification and quantities presented in Figure 1. The "Element" category encompasses specialized forms of basic visual effects elements like flames and lightning. The "Magic" category assembles imaginative magical effects, which are further subdivided into offensive effects, defensive effects, and those related to portals. The "Environment" category refers to effects that can alter the environment in a video, such as explosions and dense fog. The "Scene" category is dedicated to effects that can envelop the entire frame, altering the overall ambiance of the scene, like the frame being gradually frozen over. Lastly, the "Object" category focuses on effects that have a physical presence, such as angel wings. These videos cover the types of visual effects commonly found in the film and television industry and are sufficient for evaluating the performance of a visual effects synthesis method.

Base Videos. Our base video collection consists of 117 videos, including 71 videos sourced from public video datasets (e.g., DAVIS[3]) and other 46 videos obtained from public online platforms. These videos encompass scenes with stationary camera shots, as well as footage capturing human and animal activities. Each base video has been manually captioned, with the description length kept under

50 words to accommodate the text input limitation of fewer than 77 tokens imposed by our backbone model.

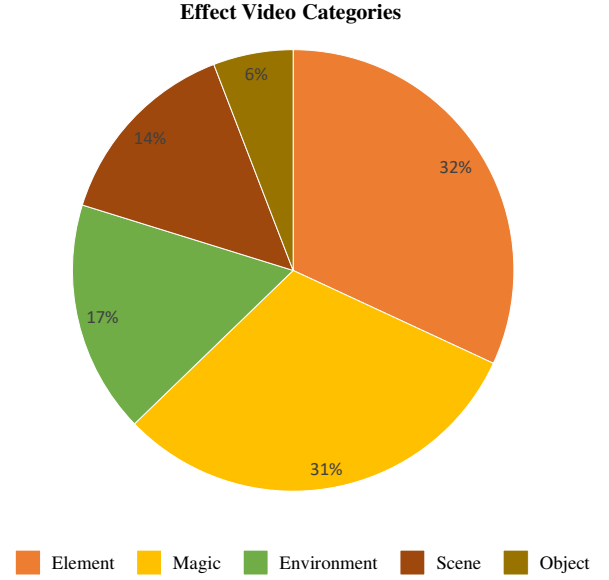


Figure 1: Statistic of effect videos categories.

3 FUTURE WORK

In the main text, we discussed the scale size and collection method of the dataset, as well as the shortcomings of the model's performance in generating rigid-body special effects, and in this section, we continue the discussion in the main text by further analyzing possible research points for AIGC-based visual effect synthesis.

Accurate Evaluation Metrics. Within the SNRE paradigm, we employ CLIP[5] score to evaluate the synthesized results from two perspectives: the similarity between the synthesized results and the original base videos, and the congruence between the synthesized results and the effect prompts. We posit that these metrics are indicative of whether the outcomes have effectively altered the base videos and generated effects that align with the textual conditions. However, we observed that for certain magical effects that are infrequently represented in CLIP's training data, such as magical shields, CLIP struggles to accurately match the results with the corresponding texts. For example, in Figure 2, The output video is more in line with the effect text than the base video, but CLIP yields the opposite score. Additionally, within the SRE paradigm, quantitatively measuring the similarity between the effects in the synthesized results and the reference effects presents an unresolved

challenge. Currently, these issues seem to be best assessed through user studies.



On a sunny day, there are two coconut trees on a beach by the ocean, and a woman is leaning against them reading a book. In the sky there is a **magic transfer door**.

Figure 2: Inaccuracy of CLIP score.

Aligned with Reference Effect and Base Video. While our method successfully integrates visual effects into base videos, the experimental outcomes reveal a loss of detail from both the base videos and reference effect videos. As shown in Figure 3, there is some loss of detail in facial features and clothing styles. This loss of detail could potentially impede the application of AIGC-based visual effects synthesis in industrial settings. In fact, preserving the original details of areas that do not require alteration is currently a focal point of research in video editing tasks. The null-text inversion technique[2] suggests optimizing unconditional textual embedding that is used for classifier-free guidance for each timestamp, thereby enhancing the high-fidelity editing of real images. FateZero[4] introduces the use of attention maps to specifically target the editing to designated areas, ensuring the consistency of the remaining areas with the original video. LOVECon[1] combines the latent of edited and original frames to preserve the structural information of the original video. These approaches could potentially be adapted for use in visual effects synthesis to achieve a more precise integration of effects.

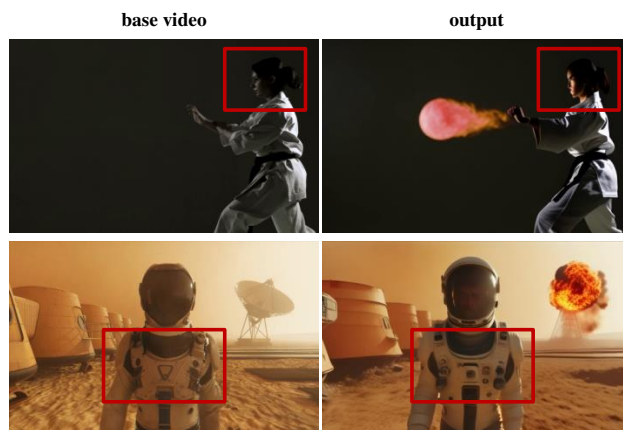


Figure 3: Loss of detail.

In summary, as an emerging application of AIGC technology, our work highlights the significant research value and application potential of visual effects synthesis tasks. However, its evaluation framework remains to be refined, and the performance of methodologies requires further enhancement. This field calls for additional research and contributions from the broader scholarly community.

4 MORE CASES

Figures 4 and Figure 5 respectively showcase the extensive experimental results of our method under the SRE and SNRE paradigms.

REFERENCES

- [1] Zhenyi Liao and Zhijie Deng. 2023. LOVECon: Text-driven Training-Free Long Video Editing with ControlNet. *CoRR* abs/2310.09711 (2023). <https://doi.org/10.48550/ARXIV.2310.09711> arXiv:2310.09711
- [2] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 6038–6047. <https://doi.org/10.1109/CVPR52729.2023.00585>
- [3] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *CoRR* abs/1704.00675 (2017). <http://arxiv.org/abs/1704.00675>
- [4] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 15886–15896. <https://doi.org/10.1109/ICCV51070.2023.01460>
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>

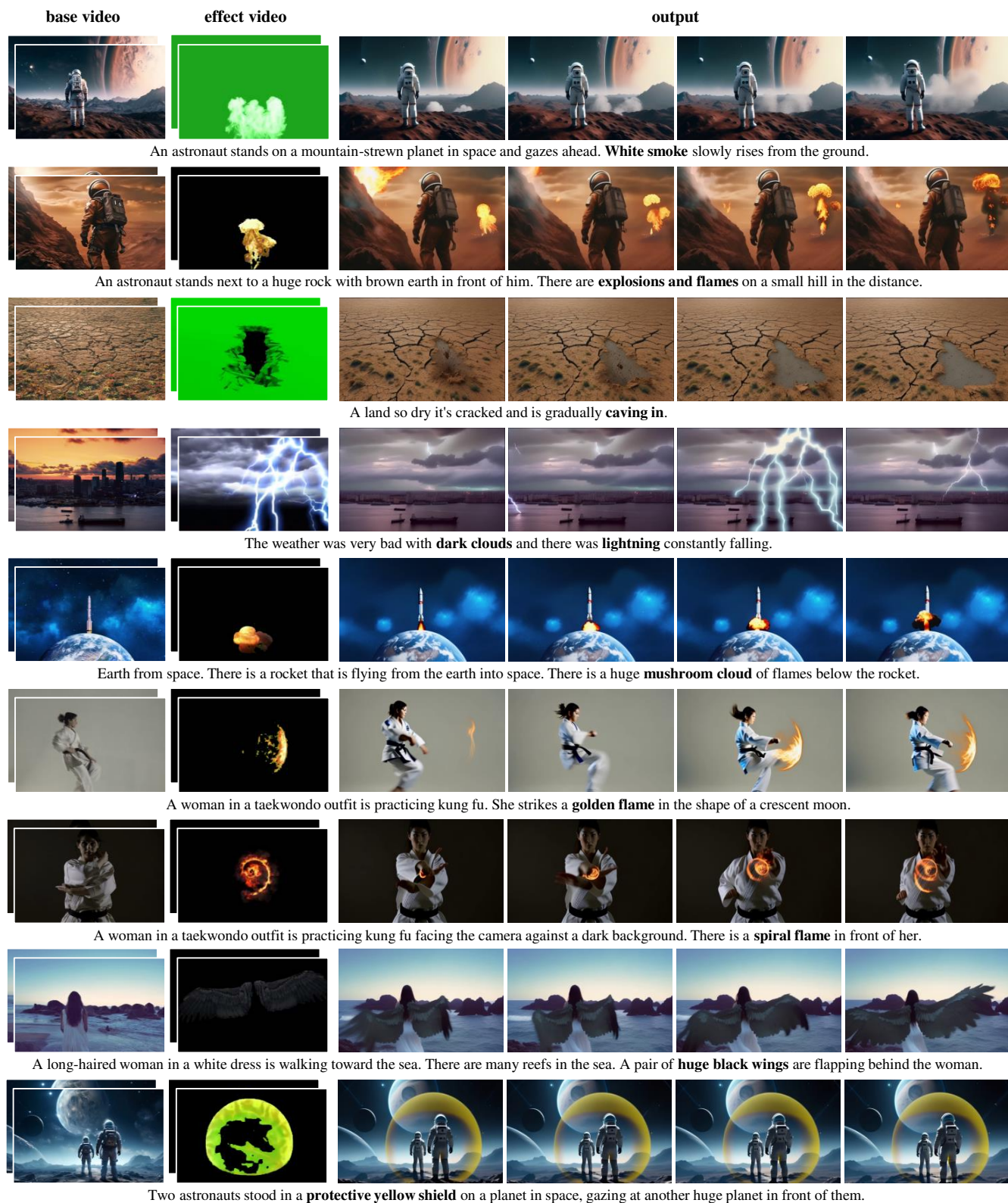
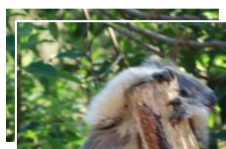
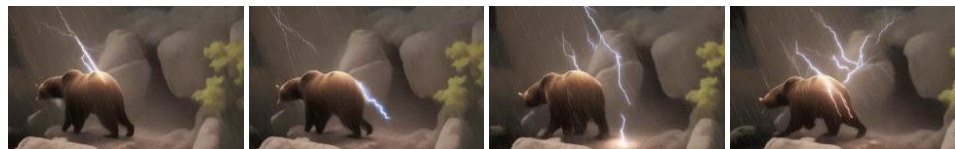


Figure 4: More cases of SRE.

base video



output



A brown bear in an enclosure walking on rocks, with **lightning** appearing all around.



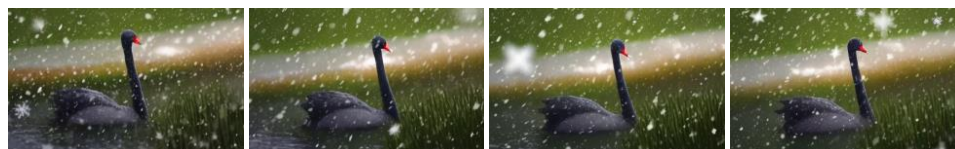
A hockey player playing hockey, when he hits the hockey, the hockey becomes a **flying fire ball**.



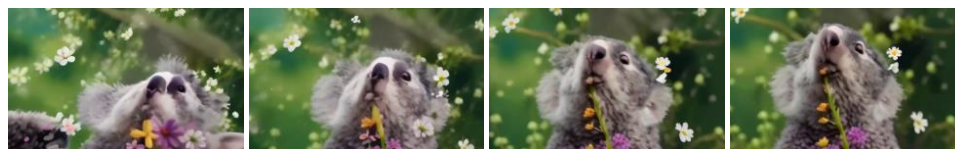
A couple riding a motorcycle surrounded by **butterflies**, on a beautiful mountain road.



A black racing car being driven down the street, surrounded by **water flow**.



A small black swan is shown swimming gracefully in a calm pond, throwing **snowflakes** backwards.



A koala climbing up a tree, surrounded by flying **flowers**.



A boy is dancing hip-hop, wearing a brown color and a pair of jeans, surrounded by **water flow**.



A **light ball** rolling through the yard of a small house surrounded by green grass and trees.

Figure 5: More cases of SNRE.