Development of mask guided visual effect generation with generative AI

IlHong Shin and Jeong-Woo Son
Media Intellectualization Research Section
Electronics and Telecommunication Research Institute
Daejeon, South Korea
ssi@etri.re.kr

***Abstract*—In this paper, we introduce a mask-guided generative AI system for controllable visual effects (VFX) video synthesis based on latent diffusion. Users can provide spatial masks to define where the effects appear and optional text prompts to describe the effect's style or semantics. The model is fine-tuned to integrate both spatial and textual conditions, enabling precise and realistic effect generation. Experimental results demonstrate that the system effectively generates high-quality, localized VFX aligned with user intent, offering a powerful tool for intuitive and automated video editing workflows.**

***Keywords*— generative AI, latent diffusion, VFX generation, spatial mask, controllable video synthesis**

I. INTRODUCTION

Recent advances in generative video modeling have enabled the creation of realistic and temporally consistent videos from various modalities such as text, images, and motion cues. Among these, Stable Video Diffusion (SVD) has demonstrated strong performance in generating diverse content-rich videos through an autoregressive latent diffusion pipeline [6]. However, existing models like SVD are primarily unconditional or limited to coarse conditioning, making them unsuitable for tasks that require precise spatial control, such as visual effects (VFX) generation in professional content workflows.

To address this limitation, we propose a mask-guided VFX generation system built upon the SVD framework. Our method introduces spatial masks and textual prompts as conditioning inputs, allowing localized and semantically consistent effects to be synthesized in designated regions of the video. This is particularly useful for post-production, animation, and AR/VR applications where users seek fine-grained control over visual elements.

By fine-tuning a pretrained SVD model with a custom dataset containing annotated VFX regions and associated prompts, our approach learns to align visual dynamics with both spatial constraints and semantic intent. The system can supports user-driven generation with controllable effects such as explosions, flames, or lighting confined to masked areas.

This paper details the system architecture, training methodology, and experimental results demonstrating the effectiveness of our approach in controllable VFX synthesis.

II. PROPOSED METHOD

The model based on Stable Video Diffusion in Fig. 1 is further fine-tuned using three conditioning inputs: spatial mask, text prompt, and latent noise. These inputs guide the generation of VFX effects within designated regions of the video. The output consists of video frames with localized visual effects consistent with the input constraints. This pipeline enables controllable and semantically aligned VFX generation.

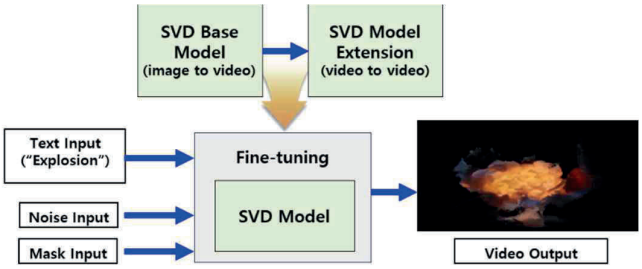

Fig. 1. Schematic representation of the proposed method

Our proposed system builds upon the Latent Diffusion Model (LDM) framework, extended to support mask-guided conditional generation of visual effects (VFX).

The architecture consists of a VAE encoder that compresses input video frames into compact latent representations. A mask encoder is used to transform binary or soft-valued spatial masks into formats compatible with the latent resolution. In parallel, a text encoder such as CLIP converts natural language prompts into semantic embeddings. These components are integrated into a U-Net-based latent diffusion model that denoises latent variables through multiple steps, while leveraging cross-attention mechanisms to guide generation with text condition.

Let z_t denote the latent noise at timestep t, m the spatial mask, and c the encoded text condition. The model is trained to predict the noise residual ε_θ using: $\boldsymbol{\varepsilon_\theta (z_t , m_c , t \mid c)}$.

To emphasize correct denoising in the masked regions, we adopt a mask-weighted diffusion loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\lambda_m \cdot \tilde{m}_{\text{vae}} \cdot \|\epsilon_\theta(z_t, t \mid m_{\text{cond}}, c) - \epsilon\|^2 + (1 - \tilde{m}_{\text{vae}}) \cdot \|\epsilon_\theta - \epsilon\|^2\right]$$

where $\varepsilon \sim N(0, I)$, $\lambda_m = 2$, m_{vae} is the mask downsampled to the latent resolution having mask value, and m_c is the VAE-encoded representation of the original mask.

To fine-tune the model, we curated a custom dataset by collecting public internet video clips featuring common VFX categories, such as explosions, fire, smoke, sparks, and light trails. Each video clip was manually or semi-automatically annotated with frame-aligned spatial masks indicating the

presence of visual effects. These masks were either binary or soft-valued to accommodate boundary transitions. Additionally, each clip was paired with a brief textual description (e.g., "explosion in the center", "red flame"), allowing the model to associate spatial patterns with semantic context.

This dataset enables the model to jointly learn the relationships between spatial layout, semantic intent, and visual appearance. During inference, a user provides a spatial mask and an optional text prompt. The model samples latent noise and denoises it iteratively, producing video frames in which the VFX appears exclusively within the masked region, aligned with the given textual description.

III. Experimental Results

The experiments were conducted using video sequences with a spatial resolution of 384×192 pixels, generating a total of 14 frames per sequence at a temporal resolution of 7 frames per second (fps). The primary visual effect evaluated in this study was explosion effect only, conditioned by the text prompt like *"explosion at the camera center"*. Each video was generated using a freeform irregular mask to localize the effect, simulating realistic scene conditions. The generation process was executed on a GPU, requiring approximately 3.2 seconds per frame (NVIDIA GTX 3080). During inference, the model utilized 24 DDIM sampling steps to progressively refine the latent noise into coherent visual outputs we want.

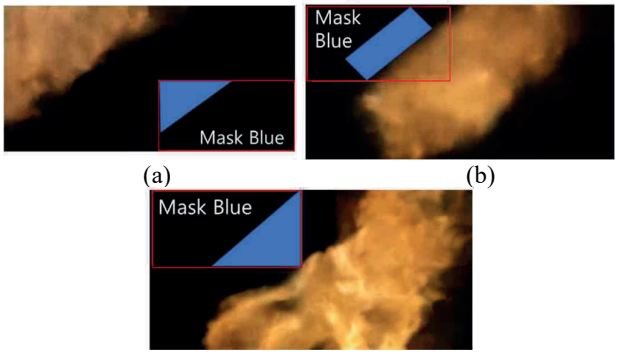

Fig. 2. Visual inspection of the generated VFX explosion effect with mask guide : (a) upper slanted mask, (b) middle slanted mask and (c) lower slanted mask

Figure 2 illustrates three representative frames from a single 14-frame video sequence generated using the proposed mask-guided VFX effect generation system. Each subfigure captures a different point in time, demonstrating how the model handles dynamically varying spatial masks for localized VFX generation. Fig.2 (a) shows a binary spatial mask applied to the top-left corner in a slanted triangular shape. This mask specifies the region for VFX placement and serves as the primary spatial condition for the diffusion model. At the midpoint of the video in Fig.2 (b), the mask evolves into a symmetric, slanted region spanning both the left and right sides of the frame. The explosion effect is generated within the masked area, but due to the stochastic nature of diffusion-based generation, some portions of the explosion may slightly extend beyond the masked boundary

This figure shows the system's capacity for temporally dynamic and spatially adaptive VFX generation, responding flexibly to user-defined masks while preserving visual coherence across time.

IV. Conclusion

In this work, we presented a mask-guided VFX generation system based on the Stable Video Diffusion (SVD) architecture. By introducing spatial masks and text prompts as conditioning signals, our model enables localized control over the placement and style of visual effects in video synthesis. We fine-tuned a pretrained SVD model on a custom dataset focusing on explosion effects, demonstrating the feasibility of spatially constrained generation using latent diffusion techniques.

While the current system supports only explosion-type effects, future work will expand the scope to include a broader range of VFX categories such as smoke, fire and watering. This will involve building a more diverse and richly annotated dataset, optimizing training pipelines for multi-class conditioning, and improving temporal coherence across frames. Additionally, we plan to integrate a user-friendly interface for interactive mask design and prompt input, enabling creative professionals to intuitively control and deploy AI-generated visual effects in real-world production environments.

ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00395401, Development of VFX creation and combination using generative AI)

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[3] A. Singer, R. Bakhtin, K. Misra, et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv preprint arXiv:2209.14792, 2022.

[4] B. Kawar, M. Elad, et al., "Imagic: Text-Based Real Image Editing with Diffusion Models," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2023.

[5] B. Zhang, Y. Shi, M. He, and Y. Duan, "MagicBrush: Interactive Image Editing with Region-Aware Diffusion Models," Proc. Int'l Conf. on Computer Vision (ICCV), 2023.

[6] J. Khachatrian, M. Zhang, H. Shi, D. Adcock, S. Vemprala, et al., "Scaling Autoregressive Models for Content-Rich Video Generation," *arXiv preprint arXiv:2312.06636*, Dec. 2023.