Supplementary Material

ANONYMOUS AUTHOR(S)

ACM Reference Format:
Anonymous Author(s). 2025. Supplementary Material. 1, 1 (January 2025), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Implementation Details

1.1 Models

We use a publicly available CogVideoX-5B [2, 6] text-to-video model, which is trained on video clips of the length of up to 49 frames and 720x480 resolution. Consequentially, our results are in the same resolution with the same number of frames. This model is a transformer-based model that processes both text and video modalities together. For text-based segmentation the prominent objects in the video and the newly generated content we utilize EVF-SAM [8] - a text-base video segmentation model based on SAM2 [4]. Our vision-language model of choice is GPT-4o [3], which we use through the provided Python API.

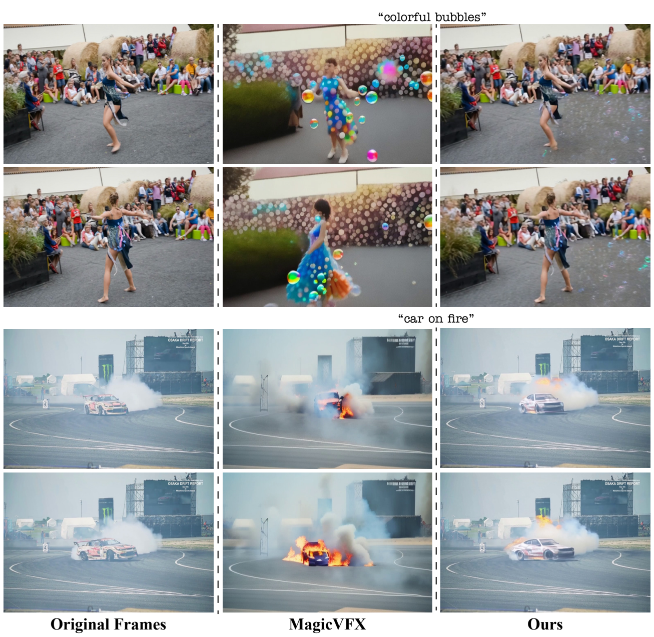

Fig. 1. Comparison to MagicVFX. The result of MagicVFX the output differs significantly from the original video.

ACM XXXX-XXXX/2025/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.2 Keys and Values Extraction

Following [5, 7], to obtain T2V diffusion model intermediate latents, we use DDIM inversion (applying DDIM sampling in reverse order) on the input video, using 1000 forward steps, with an empty string as text prompt. During the forward pass in our method, the intermediate latents are used for the extraction of keys and values.

1.3 VLM Prompting

While the model gives an accurate, descriptive source scene caption, in some cases, we observed that it fails to give captions suitable for compositing VFX with the scene. To overcome this, we ask the model to imagine a conversation with a visual effects (VFX) artist to obtain a caption that would describe the composited scene correctly. In this conversation, GPT-4o will "consult" with a VFX artist about how the new content should be integrated into the scene. Based on their discussion, it will be asked to provide a caption that describes how the added content fits into the scene. This results in text prompts that encourage the generated output video to include a natural interaction between the new content and the original environment. In this prompt, we also ask the VLM to provide a list of prominent foreground objects in the original video: O_{orig} and the object that will be added according to the edit prompt: O_{edit}. The full prompt for the VLM is shown in Figure 2.

In addition, as discussed in Sec. 5.2 we utilize the VLM for interpretable quality assessment. The full set of instructions for the VLM can be seen in Fig. 3.

1.4 Latent Mask Extraction

As discussed in Sec. 4.3, we iteratively update the residual latent x_{res} in the regions where the new content appears. This requires calculating the mask of the new content in the latent space. To do this, we first apply the segmentation model [8] to the current output of SDEdit and get the mask of the new content in RGB space. However, the VAE in the T2V diffusion model involves both spatial and temporal downsampling, making it challenging to directly map RGB pixels to their corresponding latent regions. To address this, we encode the RGB masks through the VAE-Encoder and apply clustering to partition the resulting latents into two groups, effectively producing downsampled masks that align with the latent space representation.

1.5 Runtime

Our method's two most computationally intensive parts are - DDIM inversion, which takes ~15 minutes, and iterative updates of the edit residual, which takes ~20 minutes. Importantly, DDIM inversion needs to be performed only once per video and can support multiple subsequent edits, making the process more efficient when applying various modifications to the same video content.

```
You will receive a few images of the source scene and a description of new content to be added to the scene. It is possible that you will receive a source prompt as well.
 Your task is to provide two captions based on the following steps:
Source Scene Caption:
**Note**: If a source scene prompt is provided, use it as is!
Provide a detailed description of the source scene without considering the added content.
Focus on the existing objects, environment, and actions in the scene.
Ensure the description maintains the original mood and setting.
VFX Conversation:
Imagine a conversation with a Visual Effects (VFX) artist about how the new content should be integrated into the scene.
Remember, the new content can be objects or multiple objects or effect or really anything the user provides. so be clear to explain this to the VFX artist.
The new content should interact naturally with the environment (e.g., shadows, lighting, or physical elements like grass, water, or other objects) but without altering the dynamics of the source scene.
The object must fit into the scene without affecting the original characters' behavior or actions.
The interaction between new content and foreground object must be included (e.g. object A is interacting with object B). in terms of dynamics and motion as well.
Describe how the object interacts and how it blends into the scene.
Composited Scene Caption:
Based on the conversation with the VFX artist, provide a caption that describes how the added content fits into the scene.
The caption must reflect natural interaction between the new content and the environment (e.g., lighting, shadows, physical effects), while ensuring the original dynamics remain unchanged.
The content should be aware of the surroundings, but the behavior, and flow of the original scene should remain consistent.
 The overall atmosphere might change of course due to this addition to scene.
**Output format*** -  a dictionary with keys: "source_scene_caption", "vfx_conversation", and "composited_scene_caption".
- **source_scene_caption**: source_scene_caption will be - A detailed caption of the source scene. If provided, use the given caption.
- **vfx_conversation**: A simulated conversation about how the new content should be integrated into the scene.
- **composited_scene_caption**: will be - A detailed caption of the composited scene, integrating the new content.
**Note**:The composited_scene_caption and source_scene caption must each have between 90-95 words. Extra words will be ignored.
**Note**:The vfx_conversation could be as long as required in order to succeed.
**Note**: Don't start the composited_scene_caption with - "The scene now." or "Added to the scene" "Scene has transformed",
the composited_scene_caption should be understandable to anyone that does not have access to the source_scene_caption.
And you should not simply concatenate between the source and composition.
You should have an entirely new caption that describes the essence of the integrated scene with both the source content and new content.
Don't use anything similar to "now the scene"
```

Fig. 2. VLM instructions used for generating the textual descriptions.

```
You are a helpful assistant that pays attention to context and estimates the perceptual quality of provided videos, specifically for the task of integrating new content into a given video.

I would like you to help me estimate the quality of an edited videos based on the original frames along with text descriptions.

You will be shown four grids. Each grid will be of the following type: left column will contain three frames from the original video.
The next 2 columns will each contain three frames from different video editing methods. Above each column there will be a caption (original video, 1, 2, ...).
Each method's task is to integrate the new content into the source video according to the edit prompt.

The prompt describing the original video is "{original_prompt}". The edit prompt for all of the methods is "{edit_prompt}".
Now, please conduct a perceptual quality comparison in terms of 1) alignment with the edit prompt; 2) visual quality,  3) new content harmonization and 4) dynamics

For each method provide a score from 0 to 1 for each of the five criteria with higher scores indicating better results.
Your response must include a concise description regarding the perceptual quality of each method and a score to summarize quality for each criterion while well aligning with the given description.
1) When assessing the alignment with the edit prompt, consider how well the method follows the edit prompt and if the frames contain the desired content.
If the method fails to follow the edit prompt, the score should be low.
2) For visual quality, consider how realistic the frames look - are there any visual artifacts, are the lighting and colors realistic, are the objects in the image recognizable.
3) For content harmonization - how well the content is harmonized with the scene, are the proportions of the added content correct, is the depth
and perspective of the added content consistent with the scene. Is placement of the added object physically realistic - does it look like it belongs in the scene or does it look out of place.
Are the occlusions of the added content consistent with the scene.
4) For dynamics assessment - how realistically the added object is moving relatively to the scene. Is its motion aligned with the camera motion of the original video? If the object, for example floats unrealistically or flickers, the score should be low.
```

Fig. 3. VLM evaluation protocol

2 Additional comparisons

We perform an additional qualitative comparison to MagicVFX [1]. As can be seen in Fig. 1, MagicVFX struggles to remain faithful to the original scene and has lower visual quality compared to our method.

References

[1] Jiaqi Guo, Lianli Gao, Junchen Zhu, Jiaxin Zhang, Siyang Li, and Jingkuan Song. 2024. MagicVFX: Visual Effects Synthesis in Just Minutes. In *ACM Multimedia*. https://api.semanticscholar.org/CorpusID:273642722

[2] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).

[3] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David

Algorithm 1 DynVFX Algorithm

Input:
$\mathcal{V}_{\text{orig}}, \mathcal{P}_{\text{VFX}}$ ▷ Input video & instruction prompt
τ_{A} ▷ Extended Attention threshold
Ψ ▷ Video segmentation model
VLM ▷ Vision Language model

Preprocess:
$\mathcal{P}_{\text{comp}} \leftarrow \text{VLM}[\mathcal{V}_{\text{orig}}, \mathcal{P}_{\text{VFX}}]$ ▷ Composition Prompt
$\mathcal{O}_{\text{orig}}, \mathcal{O}_{\text{edit}} \leftarrow \text{VLM}[\mathcal{V}_{\text{orig}}, \mathcal{P}_{\text{VFX}}]$ ▷ Original objects and VFX object
$M_{orig} \leftarrow \texttt{Get-Latent-Mask}(\Psi; \mathcal{V}_{\text{orig}}, \mathcal{O}_{\text{orig}})$ ▷ Extract source masks
$x_{orig} \leftarrow \texttt{Encode}[\mathcal{V}_{\text{orig}}]$ ▷ Encode video into latent space
$\mathbf{K}_{\text{orig}}, \mathbf{V}_{\text{orig}} \leftarrow \text{DDIM-Inv}[x_{orig}] \quad \forall t \in [T]$

For $t = \tilde{T}, \ldots, T_{min}$ **do**
$\boldsymbol{x_{res}} = \mathbf{0}$ ▷ initialize the residual latent
$\boldsymbol{x_{comp}} = x_{orig} + x_{res}$
if $t > \tau_A$ then $K^E, V^E \leftarrow \mathcal{F}(K_{\text{orig}} | M_{\text{orig}}), \mathcal{F}(V_{\text{orig}} | M_{\text{orig}})$
else $K^E, V^E \leftarrow \emptyset$
$\boldsymbol{\hat{x}_{comp}} \leftarrow \texttt{Sampling}[x_{comp}, \mathcal{P}_{\text{comp}}, t; \texttt{AnchorExtAttn}[K^E, V^E]]$
$\boldsymbol{\hat{\mathcal{V}}_{comp}} \leftarrow \texttt{Decode}(\boldsymbol{\hat{x}_{comp}})$ ▷ Decode latent
$\boldsymbol{M_{VFX}} \leftarrow \texttt{Get-Latent-Mask}(\Psi; \hat{\mathcal{V}}_{comp}, \mathcal{O}_{\text{edit}})$ ▷ Extract VFX masks
$\boldsymbol{x_{res}} = \boldsymbol{M_{VFX}} \cdot (\hat{x}_{comp} - x_{orig})$

$\boldsymbol{x_{comp}} = x_{orig} + x_{res}$
$\mathcal{V}_{\text{comp}} \leftarrow \texttt{Decode}[x_{comp}]$ ▷ Output video
Output: $\mathcal{V}_{\text{comp}}$

Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[4] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya K. Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Doll'ar, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *ArXiv* abs/2408.00714 (2024). https://api.semanticscholar.org/CorpusID:271601113

[5] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1921–1930.

[6] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

[7] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. 2023. Space-Time Diffusion Features for Zero-Shot Text-Driven Motion Transfer. *arXiv preprint arxiv:2311.17009* (2023).

[8] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. 2024. EVF-SAM: Early Vision-Language Fusion for Text-Prompted Segment Anything Model. (2024). arXiv:2406.20076 [cs.CV] https://arxiv.org/abs/2406.20076