SUV: Suppressing Undesired Video Content via Semantic Modulation Based on Text Embeddings
Supplementary

Xiang Lv Mingwen Shao* Lingzhuang Meng Chang Liu Yecong Wan Xinyuan Chen

Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)

lvxiang1997@126.com, smw278@126.com, lzhmeng1688@163.com, z23070142@s.upc.edu.cn, yecongwan@gmail.com, b24070011@s.upc.edu.cn

Overview

In this supplementary material, we provide more implementation details and results comparison. Specifically, **Sec. A** elaborates on the specific implementation details of the our proposed SUV method. Subsequently, **Sec. B** presents more quantitative comparisons with existing methods, demonstrating the superiority of our method in terms of alignment quality and efficiency. Then, **Sec. C** provides additional qualitative results with existing methods in qualitative aspects, demonstrating the advancement of our method in terms of visual quality, editing accuracy, and temporal consistency. Finally, more visualizations of the results of the ablation study and hyperparameter study are presented are presented in **Sec. D**.

A. More Implementation Details

Additional Setup Details. Our experiments are implemented with the Pytorch framework on a NVIDIA RTX A6000 GPU. To demonstrate our method more clearly, Algorithm 1 describes the specific framework of our method.

Algorithm 1: Algorithm of SUV.

Input: Source video $\mathcal{V}_{src}$, Source text prompt $\mathcal{P}_{src}$ and Target text prompt $\mathcal{P}_{edit}$.

$\mathcal{C}_{src}$ = Text Encoder ($\mathcal{P}_{src}$),
z_0 = Image Encoder ($\mathcal{V}_{src}$) ;
DDIM inversion for latents z_0:
for $t = 1, 2, ..., T$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(z_{t-1}, t, \mathcal{P}_{src})$,
  $z_t = \sqrt{\alpha_t}\frac{z_{t-1}-\sqrt{1-\alpha_{t-1}}\epsilon_{t-1}}{\sqrt{\alpha_{t-1}}} + \sqrt{1-\alpha_t}\epsilon_t$.
end
$\hat{z_T} = z_T$.
Dividing $\hat{z}_t$ into m video groups $\{V_1, \cdots, V_m\}$;
$\mathcal{C}$ = Text Encoder ($\mathcal{P}_{edit}$), $\hat{\mathcal{C}}$ = ES-Operato ($\mathcal{C}$);
for $t = T, ..., 1$ **do**
  $\hat{\mathcal{C}}_t = \hat{\mathcal{C}}$,
  $\mathcal{X}_{FF}, \hat{\mathcal{X}}_{FF} \leftarrow$ Fusion similar features,
  $\mathcal{X}_{FD}, \hat{\mathcal{X}}_{FD} \leftarrow$ Decomposition similar features,
  A_t^{PE}, A_t^{NE} = CA ($\mathcal{X}_{FD}, \mathcal{C}_t$),
  $\hat{A}_t^{PE}, \hat{A}_t^{NE}$ = CA ($\hat{\mathcal{X}}_{FD}, \hat{\mathcal{C}}_t$),
  $\mathcal{L}_{pr} = \left\|A_t^{PE} - \hat{A}_t^{PE}\right\|^2$,
  $\mathcal{L}_{ns} = -\left\|A_t^{NE} - \overline{A}_t^{NE}\right\|^2$,
  $\hat{\mathcal{C}}_t = \text{argmin}(\mathcal{L}_{pr} + \mathcal{L}_{ns})$,
  $\hat{z}_{t-1}$ = Denoising ($\hat{z}_t, t, \hat{\mathcal{C}}_t$).
end

Output: Edited Video $\mathcal{V}_{edit}$ = Image Decoder($\hat{z}_0$).

Evaluation Metrics. The automatic metrics are based on pre-trained CLIP models. Specifically, *Temporal Consistency* evaluates the temporal consistency of the edited frames by calculating the cosine similarity between successive frame pairs. *Frame Accuracy* measures the editing accuracy for each frame, i.e., whether the CLIP similarity between the edited image and the target prompt is higher than that with the source image. Meanwhile, to further assess text-image alignment quality, we adopt *PickScore* [4], which quantifies how well a generated image semantically matches the input text prompt. The user study includes four metrics: *Editing Accuracy*, *Aesthetics Quality*, *Temporal Consistency*, and *Overall Impression*, which are used to assess the editing accuracy, aesthetic quality, temporal consistency, and overall impression of the edited video, respectively. For a fair comparison, we invite 31 subjects to

*Corresponding author

Method	Control A Video	ControlVideo	FateZero	FLATTEN	TokenFlow	Ours
PickScore	0.132	0.145	0.169	0.148	0.165	0.240

Table 1. Quantitative evaluation with baselines on PickScore. The color of each cell shows the best and the second best.

Method	Control A Video	ControlVideo	TokenFlow	Ours
Running Time / GPU Memory	8.3min / 13.17G	**4.4min** / 11.81G	6.9min / **11.24G**	8.6min / 12.07G
Temporal Consistency / Frame Accuracy	0.9751 / 0.6329	0.9809 / 0.7278	0.9824 / 0.6361	**0.9896 / 0.8681**

Table 2. Comparison of efficiency and performance.

Figure 1. **The visual comparison of our SUV and existing baselines.** Compared to the other methods, our SUV can effectively suppress the undesired content of video while maintaining overall temporal consistency of the generated video.

score the videos with different methods.

Comparison Baselines. We compared our method with five state-of-the-art video editing methods. **(1) Control A Video** [1] integrates motion priors and content priors into video generation to improve video temporary consistency. **(2) ControlVideo** [7] integrates temporarily extended ControlNet into the T2I diffusion model and utilizes information such as depth and edge maps of the original video to control the editing results. **(3) FateZero** [5] achieves first zero-shot video editing through DDIM Inversion and attention blending techniques. **(4) FLATTEN** [2] leverages an existing optical flow detection model for more accurate optical flow-guided attention learning. **(5) TokenFlow** [3] introduces a linear combination of diffusion features to enhance the consistency of the video for reducing the inter-frame flickering.

B. More Quantitative Comparisons

We report the PickScore on 15 text-video pairs in Table 1, where our method significantly outperforms existing state-of-the-art approaches. In addition, to further illustrate the efficiency of our method, we present the Running Time and GPU Memory of different methods in Table 2. It can be seen that our method strikes a balance between performance and efficiency.

C. More Visual Results

In order to more intuitively illustrate the effectiveness of our method, we conducted extensive experiments compar-

Figure 2. **The visual comparison of our SUV and existing baselines.** Compared to the other methods, our SUV can effectively suppress the undesired content of video while maintaining overall temporal consistency of the generated video.

Figure 3. **The visual comparison of our SUV and existing baselines.** Compared to the other methods, our SUV can effectively suppress the undesired content of video while maintaining overall temporal consistency of the generated video.

ing our method with other existing methods. As shown in Figure 1, Figure 2, Figure 3 and Figure 4, the results indicate that these methods struggle to accurately understand negative text prompt (e.g. "without glasses and without an earring"), resulting in undesired information still appearing in the edited video. In contrast, our approach not only ac-

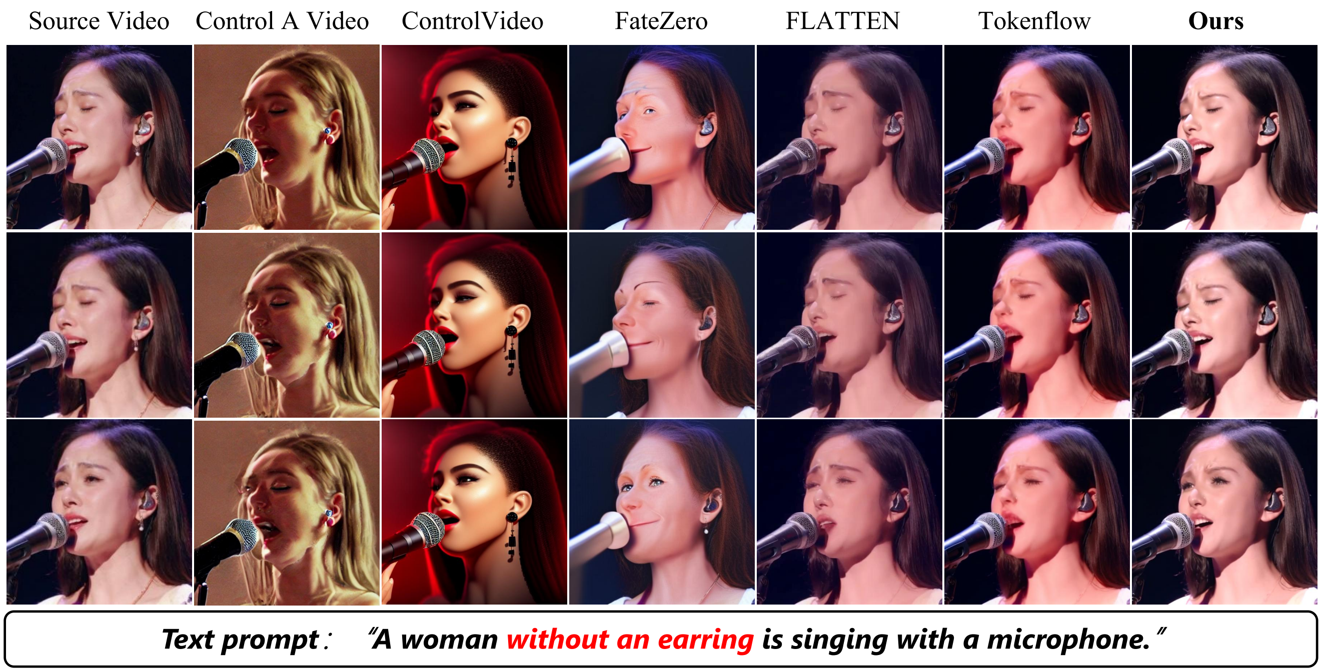

Figure 4. **The visual comparison of our SUV and existing baselines.** Compared to the other methods, our SUV can effectively suppress the undesired content of video while maintaining overall temporal consistency of the generated video.

Figure 5. **The visual comparison of our SUV and image-based editing method InfEdit.** Compared to the InfEdit, our SUV can effectively suppress the undesired content of video while maintaining overall temporal consistency of the generated video.

Figure 6. **The visual results of our method.** Our proposed SUV not only supports undesired content suppression, but also enables to realize general editing tasks such as editing grassland to snow.

curately achieves video content suppression, but also effectively maintains the temporal consistency of the video during the editing process.

Furthermore, to validate the capability of FFS in maintaining temporal consistency, we integrate it into the image-based editing method InfEdit [6] for comparison, and the results are illustrated in Figure 5. It can be seen that although InfEdit performs well in terms of temporal consistency, it fails to effectively suppress unwanted content, while our method yields superior results. Additionally, our method also supports general editing task, such as background from grassland to snow, as shown in Figure 6.

Number of m	$m = 2$	$m = 4$	$m = 6$
Temporal Consistency / Frame Accuracy	0.9689 / 0.8654	0.9748 / 0.8934	0.9735 / 0.8763

Table 3. Ablation studies on the number of video groups. The color of each cell shows the best and the second best.

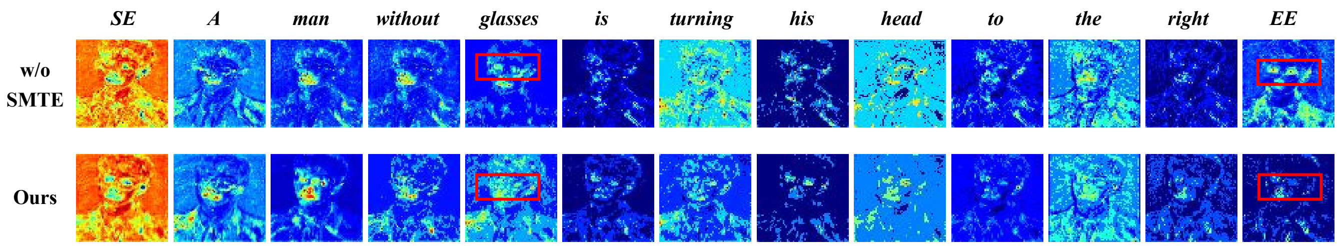

Figure 7. **The impact of the semantic modulation based on text embeddings.** It can be seen that after removing the module, the content of glass appears in both "glass" and "EE" of the text embeddings, and our SUV can effectively remove these undesired negative content.

D. More Ablation Results

We perform visual validation of semantic modulation based on text embeddings, and the results are shown in Figure 7. When removing the SMTE, we obtain each cross-attention maps corresponding to the text embeddings. It can be seen the content of glasses appear in both "glass" and "EE" of the text embeddings after removing the SMTE, it indicates the SUV can effectively remove these undesired content of video.

In addition, we also conduct ablation studies on the hyperparameter m, and the results are shown in Table 3. We adopt $m = 4$ by default as it yields the best performance.

References

[1] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2

[2] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *International Conference on Learning Representations*, 2024. 2

[3] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *International Conference on Learning Representations*, 2024. 2

[4] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 1

[5] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 2

[6] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. 2024. 4

[7] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *International Conference on Learning Representations*, 2024. 2