Preserving Source Video Realism: High-Fidelity Face Swapping for Cinematic Quality

Supplementary Material

Considering the space constraints of the main paper, this supplementary material provides additional experimental results and presents the construction details of *Face2Face* and *CineFaceBench*. The content is organized as follows:

- Sec. A: Generalization Beyond Train Data Quality.
- Sec. B: Keyframe Identity Injection for Accumulated Identity Errors.
- Sec. C: Robustness to Keyframe Quality.
- Sec. D: Potential of Keyframe Selection.
- Sec. E: Robustness to Identity Differences.
- Sec. F: Robustness to Attribute Variations in Source Video.
- Sec. G: Grayscale Keyframe Guidance for Robust Color Learning.
- Sec. H: *Face2Face* Construction Details.
- Sec. I: *CineFaceBench* Construction Details.
- Sec. J: Comparison with Closed-Source Methods.
- Sec. K: Limitations of LIVINGSWAP.

A. Generalization Beyond Train Data Quality

In Sec. 5.3, we analyzed the robustness of LIVINGSWAP under varying levels of data quality, demonstrating the robustness of our model to failed noisy train data. To further investigate whether LIVINGSWAP is fundamentally constrained by the quality of the training dataset itself, we conduct an additional experiment. We manually select several noisy source–target pairs from *Face2Face*. These pairs contain various types of degradation, including local failure cases caused by failed swaps (e.g., residual beards), artifacts and misaligned expressions arising from large identity gaps or occlusions, as shown in Fig. 5. We then run LIVINGSWAP on these noisy pairs using the exact same inference process as Inswapper, which was used to construct the dataset.

As illustrated in Fig. 6, LIVINGSWAP consistently surpasses the quality of the original dataset pairs, producing results with improved expression alignment, visual realism, and significantly fewer local artifacts. We attribute this improvement to two key design choices. (1) Reversing the role of data when constructing training pairs, which ensures reliable ground-truth supervision even when the original swaps contain noise. (2) Strong priors in the pretrained model, which enable the system to robustly correct misaligned or corrupted supervision. Together, these factors allow LIVINGSWAP to generalize beyond the limitations of the training data and deliver high-quality, noise-resistant face swapping results.

Table 5. The potential of rule-based keyframe selection.

Methods	ID Sim. ↑	Expr. ↓	Light ↓	Gaze ↑	Pose ↓
Inswapper	0.224	2.609	0.248	0.587	3.867
LivingSwap (Fixed-interval Keyframe)	0.073	2.441	0.243	0.634	**3.240**
LivingSwap (Rule-based Keyframe)	**0.240**	**2.393**	**0.236**	**0.681**	3.504

B. Keyframe Identity Injection for Accumulated Identity Errors

As demonstrated in Sec. 5.3, keyframes play a crucial role in maintaining identity consistency within the video reference paradigm. The design of keyframes not only mitigates the interference caused by the source video's identity but also plays a pivotal role in resolving accumulated ID errors. As shown in Fig. 7, when using only the first frame as guidance and combining it with temporal stitching (as detailed in Sec. 4.3) for long video generation, ID errors gradually accumulate as the video progresses, eventually leading to significant deviations from the target face.

In contrast, by injecting keyframe identities alongside temporal stitching, our method ensures a smooth connection with the previous video chunk while also providing correct ID guidance at the end of each chunk, preventing the accumulation of errors. This entire process is illustrated in the temporal stitching part of Fig. 3.

C. Robustness to Keyframe Quality

To examine the sensitivity of LIVINGSWAP to keyframe quality, we employ various image-level face swapping models as the Per-frame Edit module. As shown in Fig. 8, LIVINGSWAP demonstrates strong robustness against degraded or inconsistent keyframes. Even when the injected keyframes contain artifacts, expression misalignment, our method still produces results that remain well aligned with the source video and perceptually more faithful. We attribute this robustness to two key factors: (1) Directly referencing the source video, which enables the model to correct erroneous keyframe guidance by restoring the appropriate visual attributes; and (2) The strong generative prior of diffusion models, which further enforces temporal realism and semantic consistency throughout the video.

D. Potential of Keyframe Selection

Although LIVINGSWAP demonstrates good robustness when handling keyframes with low face-swapping quality, using high-quality keyframes can significantly improve the overall video quality. Due to the limitations of current face-swapping models in handling profiles or extreme angles, we

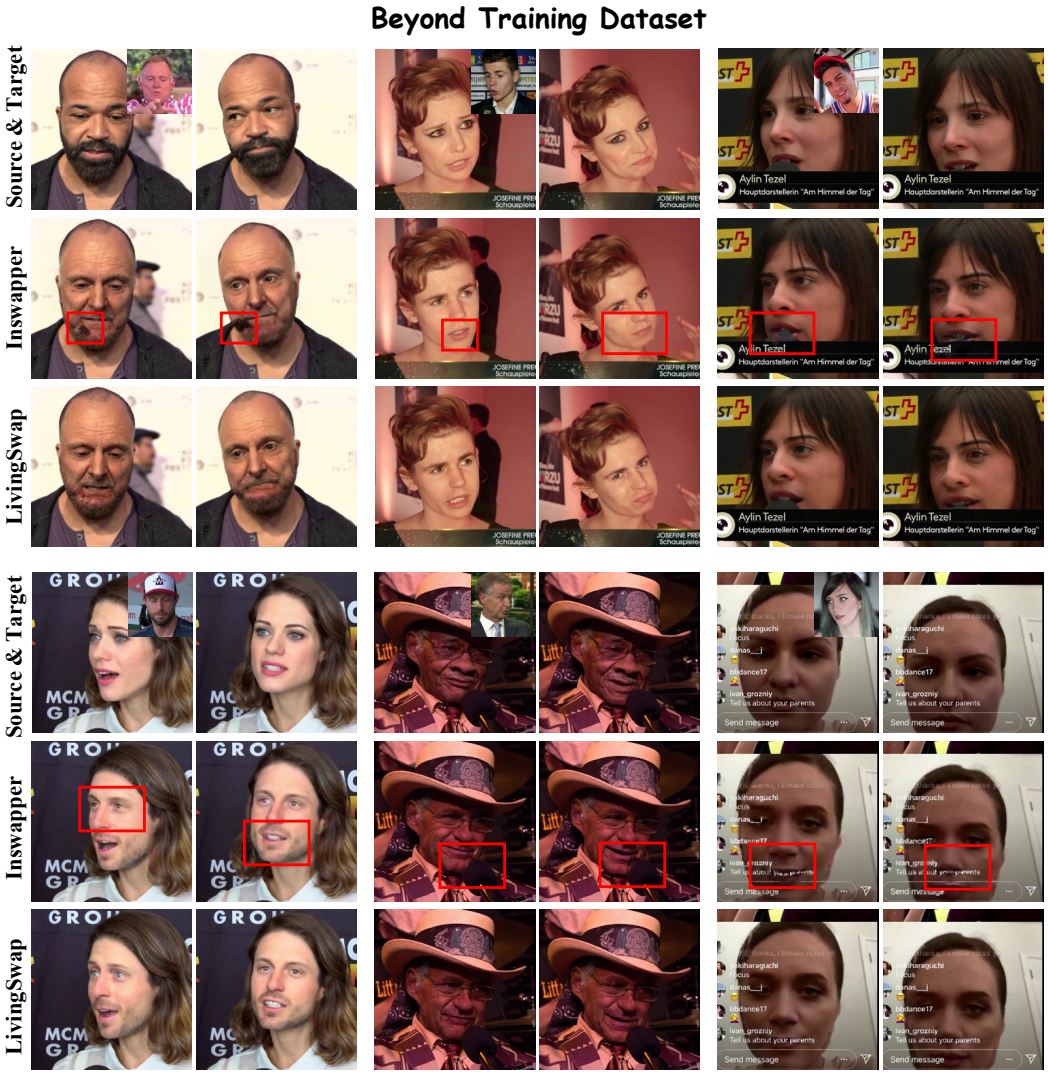

Figure 6. Qualitative comparison between the data pairs in *Face2Face* (by Inswapper [15]) and corresponding results generated by LIVINGSWAP. Benefiting from reversing the role in data pair and strong priors in pretrained model, LIVINGSWAP surpasses the quality of its training data, achieving better expression consistency and overall realism. Unlike Inswapper-based results, our method avoids local failure cases—such as incomplete swaps, mismatched regions, and occlusion-induced artifacts—demonstrating its strong generalization beyond the training dataset.

introduce a simple yet effective keyframe selection method: selecting frontal frames as keyframes based on face orientation detection (yaw within ±30° and pitch within ±20°). For evaluation, we selected the 10 worst-performing cases from the *CineFaceBench* for comparison. As shown in Tab. 5, using this straightforward keyframe selection rule greatly enhances the quality of the face-swapped video results.

Furthermore, by leveraging the keyframe guidance feature of LIVINGSWAP, we can manually refine the face-swapping results or perform post-processing modifications (e.g., adjusting appearance or makeup) using tools like PhotoShop, providing greater flexibility in enhancing and fine-tuning the final face-swapping outcomes.

E. Robustness to Identity Differences

For the scenario of swapping different identities for the same source video, we conducted experiments with multiple videos and identities. As shown in Fig. 9, leveraging the advantages of keyframe identity injection, LIVINGSWAP achieves satisfactory results for the same video, regardless of whether the identity difference is large or small. We hypothesize that this robustness of identity difference is due to the diversity of identities in our training data, as discussed in Sec. 5.3.

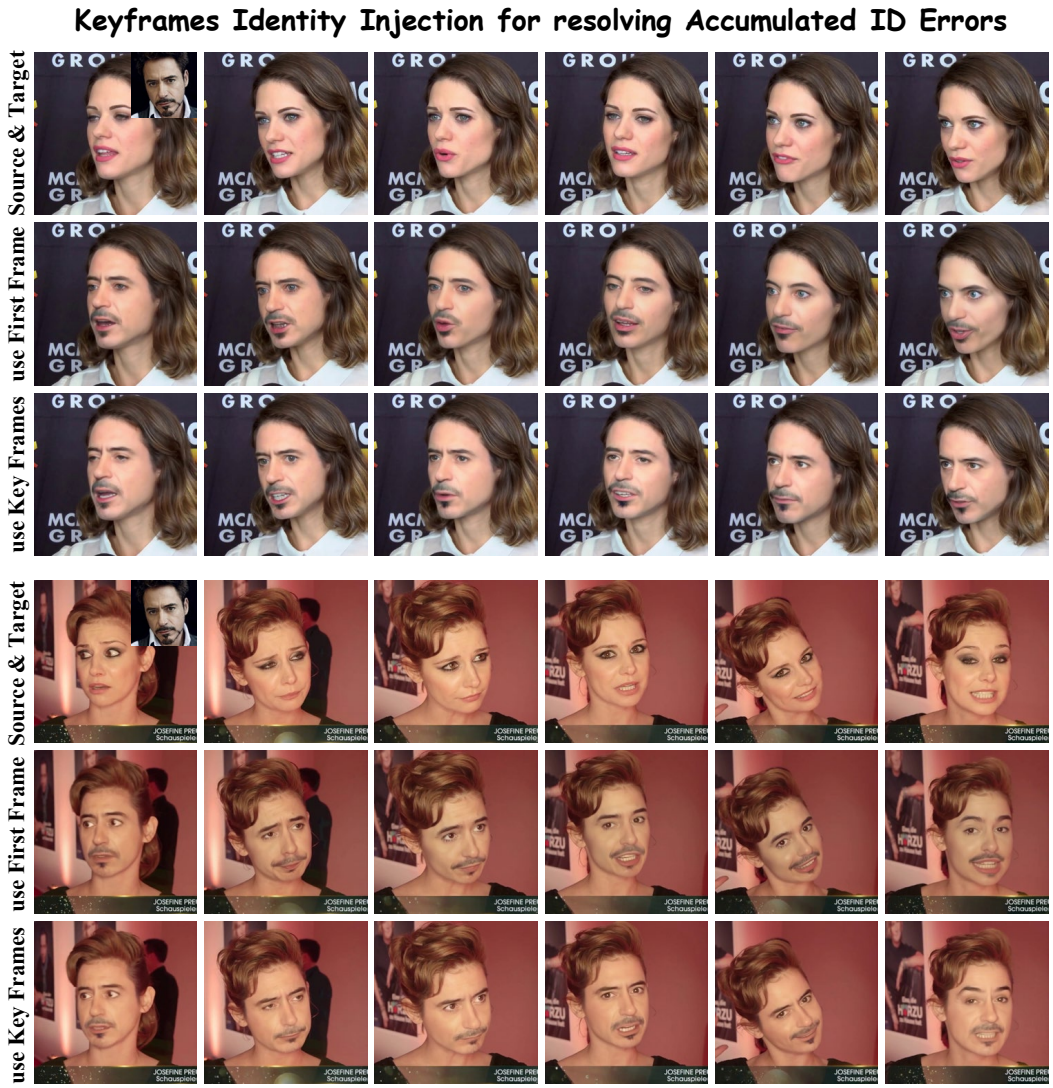

Figure 7. Keyframes Identity Injection for resolving Accumulated ID Errors. When using only the first frame for ID injection, face-swapping results suffer from gradually accumulating ID errors over time. In contrast, with Keyframe Identity Injection, each video chunk is corrected individually by swapped keyframe, ensuring better ID consistency throughout the entire long video sequence.

F. Robustness to Attribute Variations in Source Video

To verify whether our reference-based video face swapping approach is robust to attribute variations in the source video, we selected a diverse set of videos as source inputs and conducted experiments using the same target identity. As shown in Fig. 10, our model consistently produces high-quality results across attributes in challenging scenarios, such as occlusions, side profiles, and complex lighting conditions. Furthermore, owing to the robustness of keyframe quality, our model is able to generate realistic, high-fidelity outputs even when the keyframe model produces suboptimal results.

G. Grayscale Keyframe Guidance for Robust Color Learning

As shown in Fig. 12, we observe that grayscale keyframe guidance significantly reduces color bleeding and flickering artifacts in challenging cases where the keyframe edits exhibit inconsistent or unrealistic colors, while maintaining comparable identity preservation and temporal consistency.

In the main paper, LIVINGSWAP relies on RGB keyframes as temporal anchors for identity injection (Fig. 3). While this design is effective for propagating identity information, we observe a failure mode when the per-frame edited keyframes contain imperfect color statistics, e.g., incorrect skin tone or illumination caused by upstream

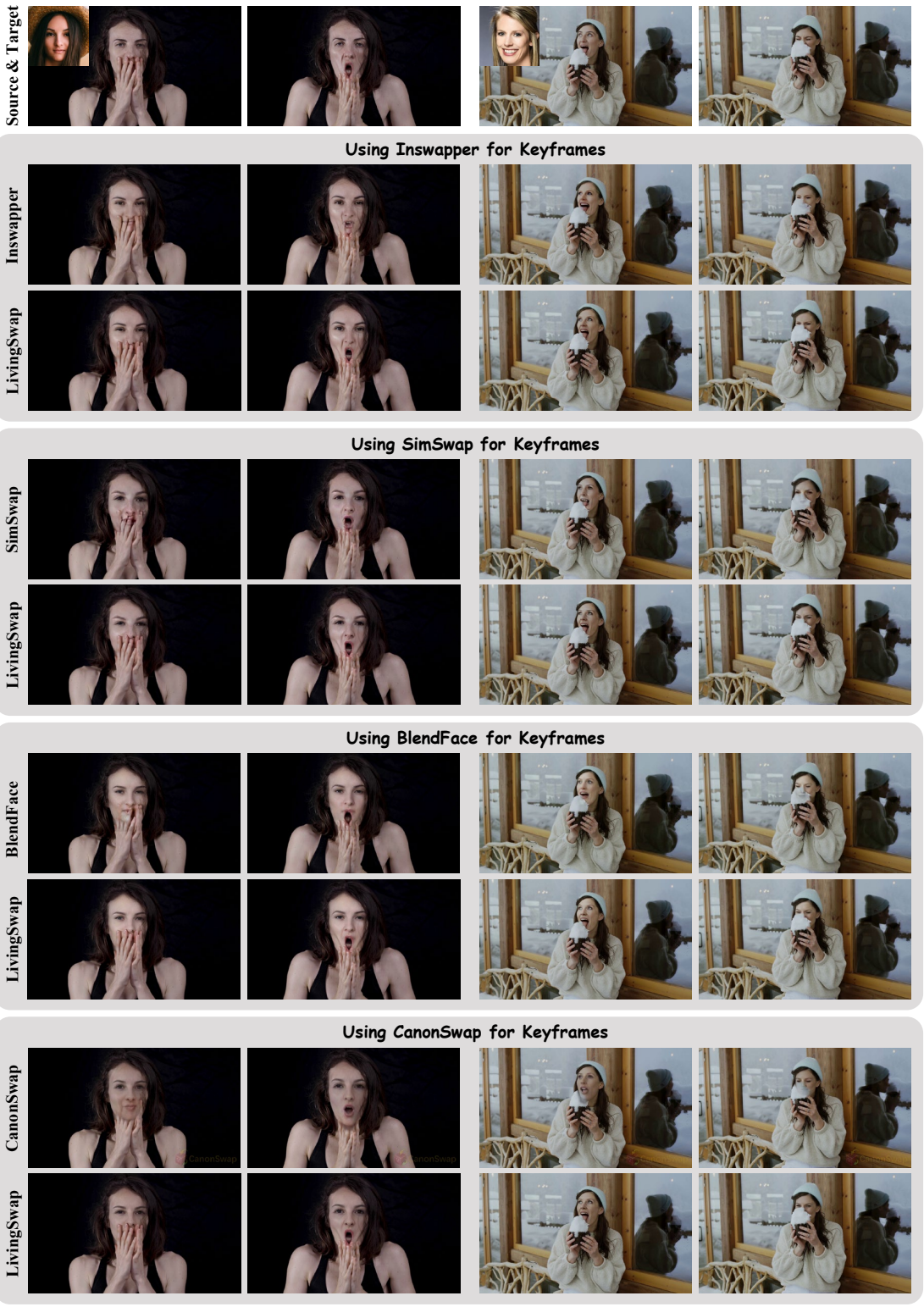

Figure 8. Qualitative comparison of using different image-level face swapping models as Per-frame Edit module. Injected keyframes often exhibit flaws including artifacts and expression misalignment. In contrast, by directly referencing the source video, LIVINGSWAP successfully refines these flaws using the corresponding source attributes, demonstrating strong robustness to imperfect or corrupted keyframes.

Figure 9. Identity swapping results on the same source video with different target identities. Our method produces consistent and high-fidelity face swaps regardless of large or small identity differences, demonstrating strong robustness to identity variations.

editing tools, as shown in Fig. 12. Because the keyframe tokens are directly concatenated with the video tokens, such color biases can be mistakenly treated as a strong supervision signal, leading the diffusion model to reproduce the wrong colors in all synthesized frames.

To mitigate this issue, we introduce a simple yet effective modification: *grayscale keyframe guidance*. As shown in Fig. 11, given an edited keyframe, we convert it to a single-channel luminance image, and then replicate this channel to form a three-channel grayscale keyframe before feeding it into the VAE encoder. The rest of the pipeline, including token concatenation and the DiT-based video generation, remains unchanged. This modification removes explicit chromatic information from the keyframe while preserving high-frequency structural cues such as facial identity, hairstyle, and local shading.

Intuitively, grayscale keyframes encourage the model to use the keyframe primarily as a structural and temporal anchor for stable identity injection, and to recover color statistics from the reference video branch in the Video Reference Completion module for fidelity. As a result, LIVINGSWAP becomes less sensitive to color artifacts in the keyframe edits.

We finetune the final LivingSwap checkpoint for 5,000 steps to adapt it to grayscale pipeline. As shown in Fig. 12, we observe that grayscale keyframe guidance significantly reduces color bleeding and flickering artifacts in challenging cases where the keyframe edits exhibit inconsistent or unrealistic colors, while maintaining comparable identity preservation and temporal consistency.

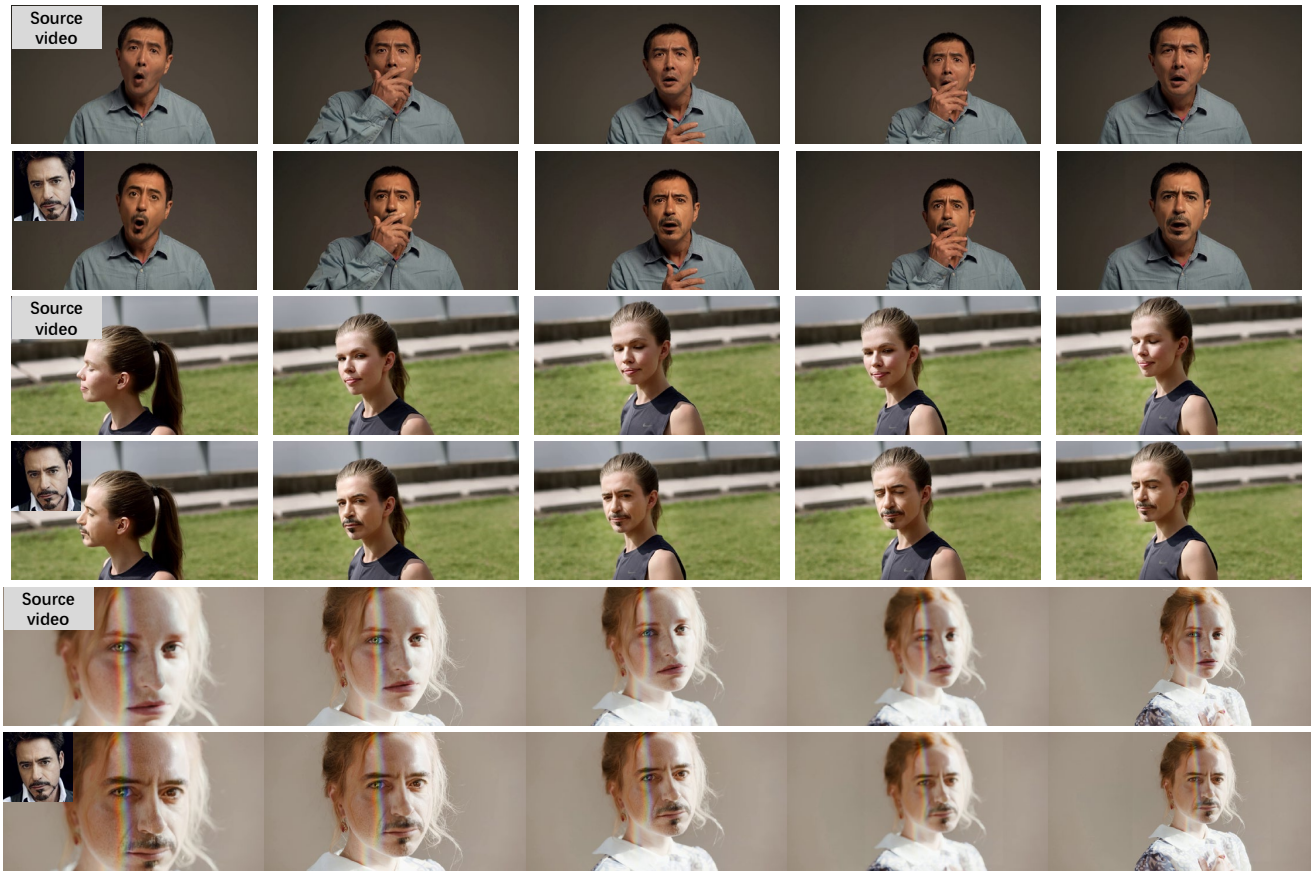

Figure 10. Face swapping results on diverse source videos with the same target identity. Our method consistently preserves target identity and produces high-fidelity outputs across challenging conditions, including occlusions, side profiles, and complex lighting.

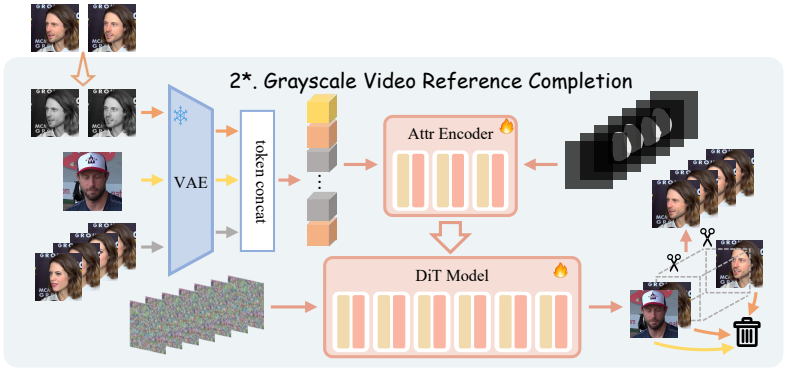

Figure 11. **Grayscale keyframe guidance.** To avoid incorrect color propagation from imperfect edited keyframes, we modify Video Reference Completion module and convert each keyframe to a grayscale image before VAE encoding. This preserves structural cues (identity, pose, shading) while removing misleading chromatic information, allowing the model to recover accurate colors from the reference video.

H. *Face2Face* Construction Details

We construct our dataset *Face2Face* based on CelebV-Text [47] and VFHQ [41]. First, we perform crop, resize, and clipping operations on the dataset to ensure the resolution is 640×640 pixels and the video length is approximately 200 frames. We then randomly pair the data and extract the first frame from the target video as the target face image. Next, we apply Inswapper [15] to perform face-swapping on the entire dataset. The process is conducted using 8

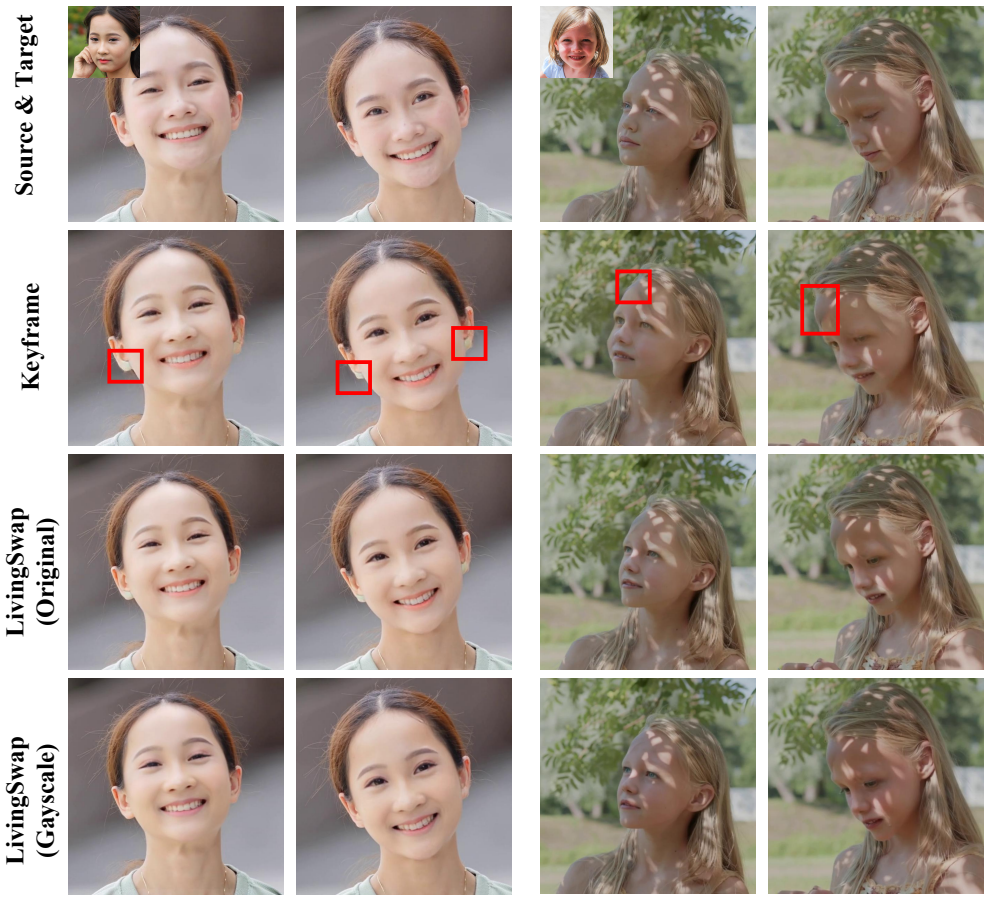

Figure 12. Compared with the original LIVINGSWAP, using grayscale keyframes effectively suppresses color bleeding (e.g., the blue tint near the ear in the first example) and reduces temporal flickering artifacts (e.g., the dark patches on the head in the second example), leading to more stable and faithful video face swapping results.

NVIDIA H100 GPUs over a duration of 120 hours. Additionally, we use the face-parsing model [46] to generate the face mask video. For the ablation study on the inpainting paradigm, we also use the pose estimation model [45] to generate the corresponding pose video. After filtering out the failed samples from the processing steps, our dataset *Face2Face* contains a total of 152,221 video samples, with a cumulative duration exceeding 300 hours. Finally, we reverse the roles in each training pair: the swapped video is used as the source video (model input), while the original video serves as the target video (ground-truth supervision).

I. *CineFaceBench* Construction Details

Due to the fact that most scenes in FaceForensics++ [30] consist of relatively simple settings such as interviews, hosts, or live broadcasts, it does not evaluate critical aspects often required in film scenarios, such as the model's ability to preserve facial expressions, lighting, makeup, and overall fidelity after face-swapping, as well as the stability of long video clips longer than 30 seconds. To address the limitations of the aforementioned benchmark, we have constructed a film scene face-swapping benchmark, *CineFaceBench*.

CineFaceBench consists of 200 video clips, paired with easy and hard target face images, resulting in 400 data pairs. Specifically, we downloaded and selected 100 video clips from two free video websites, Pixabay [27] and Pexels [26]. Additionally, we selected 100 video clips from the OpenHumanVid dataset [23] and preprocessed them, resulting in 200 video samples used for evaluation. As shown in Fig. 13, these 200 clips include challenging examples from film scenes, featuring difficult scenarios such as unique lighting, exaggerated expressions, micro-expressions, special makeup, occlusions, and even facial deformations. In addition, there are several video cases that are longer than 1 minute.

On the other hand, we randomly selected 1,000 faces from the FFHQ dataset as target face images. By calcu-

lating the ID similarity (refer to Sec. 5.1) between these faces and the source video, we selected two samples with the most similar and least similar IDs to the source video, representing the easy and hard cases, respectively. This setup allows for a better evaluation of the model's robustness to ID differences. As shown in the Tab. 1, Fig. 4 and Fig. 13, our model demonstrates impressive performance in the challenging film scene scenarios.

J. Comparison with Closed-Source Methods

Recently, several inpainting-based video face swapping methods using the Stable Video Diffusion model [2] are proposed, such as HiFiVFS [5] and FaceAdapter [14]. However, these methods are not open-source. To enable a comparison with them, we captured several demos from their project websites and conducted tests using the same target face image. The comparative results are shown in the Fig. 14. Our approach better preserves the original video attributes such as lighting and expression, and also demonstrates strong stability in occluded cases.

K. Limitations

LIVINGSWAP achieves better fidelity by directly referencing the source video, while enabling more flexible identity control and improved stability in long videos through keyframe identity injection. However, this framework design also introduces certain limitations: 1) Dependence on Keyframe Quality: The keyframe identity injection method creates a reliance on the quality of the selected keyframes. Although we demonstrated in Sec. C that LIVINGSWAP shows strong robustness to keyframe quality, it still encounters issues when dealing with keyframes that yield poor face-swapping results (e.g., identity or expression drift, image distortion, etc.). In such cases, the final output can be biased by the keyframe. This is supported by the experimental results in Sec. D, where selecting higher-quality keyframes led to significant improvements on the 10 worst-performing cases. 2) Slow Inference Speed: The use of a large video dataset to train the DiT model, combined with the Attribute Encoder for condition injection, significantly impacts the face-swapping speed of LIVINGSWAP. In our experiments, for a video of 81 frames (approximately 3 seconds at 25 fps), LIVINGSWAP requires 195 seconds (about 3 minutes) on a single H200 GPU with 108 GB memory. This translates to approximately 1 minute per second of swapped video. In conclusion, our future work will focus on exploring better methods for identity injection and optimizing the face-swapping speed.

Figure 13. Additional Qualitative Comparison of Different Methods on *CineFaceBench*. LIVINGSWAP produces results with higher fidelity and realism compared to other methods.

Figure 14. Qualitative comparison with recent inpainting-based video face swapping methods [5, 14] shows that our approach better preserves source video attributes (e.g., lighting and expression) and achieves greater stability under occlusions.