V-PartSwap: Motion-Consistent Facial Part Transfer in Videos via Alignment-Aware Diffusion

Anonymous CVPR VGBE submission

Paper ID *****

Abstract

Facial part transfer in videos, such as swapping eyes, eyebrows, nose, or mouth between identities, remains underexplored compared to full-face video swapping. While image-based methods exist, extending this task to videos is challenging: the transferred part must follow the target's motion and expression while remaining visually consistent with surrounding regions. Directly injecting reference appearance often leads to pasted artifacts or temporal inconsistencies. We propose V-PartSwap, a training-free framework for reference-guided facial part transfer built upon Diffusion Autoencoders (DiffAE). Our approach performs region-restricted semantic feature blending in the DiffAE encoder to inject reference appearance while preserving non-edited regions. To ensure motion consistency, we align reference parts to the target video using landmark-driven thin-plate spline (TPS) warping and enforce temporal coherence through region-guided flow-based attention. We also introduce a new benchmark for facial part transfer in videos. Experiments show that our method produces motion-consistent edits with competitive visual quality compared to existing editing methods.

1. Introduction

Part-level facial editing is increasingly demanded in digital media production. In particular, facial part transfer in videos enables applications such as avatar customization and synthetic dataset creation, where users modify a specific region (e.g., eyes or mouth) while preserving the person's overall identity, expression, and motion (Figure 1). Unlike full-face video swapping, this task requires introducing identity cues locally while maintaining compatibility with the surrounding facial appearance and motion.

However, facial part transfer in videos remains challenging. Facial regions are strongly coupled with their surroundings in both appearance and motion. Existing image-based methods, *e.g.* FuseAnyPart [20], mainly focus on spatial realism and treat frames independently, which often leads to temporally inconsistent edits when applied to videos. Simply injecting reference appearance without considering geometric and motion compatibility can produce pasted-looking artifacts or unstable temporal behavior.

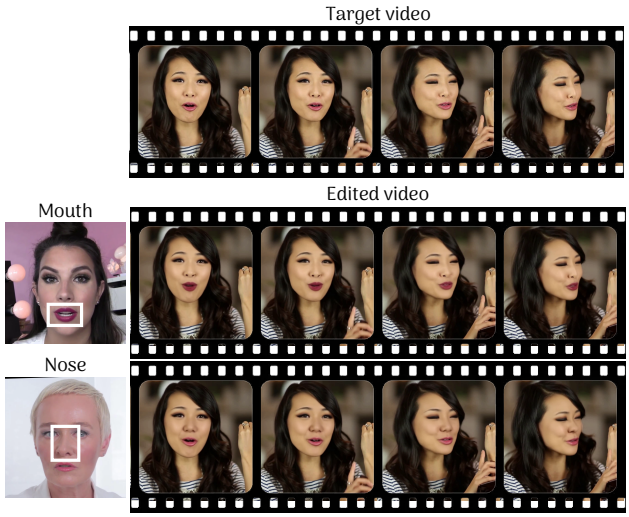

Figure 1. Given a target video (top) and reference images specifying the mouth or nose (left), our method generates edited videos (right) in which the transferred part follows the target's motion while maintaining a realistic appearance and temporal consistency.

We propose V-PartSwap, an alignment-aware framework for facial part transfer in videos built upon a reconstruction-based diffusion autoencoder (DiffAE) [15]. Our goal is motion-consistent editing, where the transferred appearance follows the target video's motion while remaining visually consistent with surrounding regions. To achieve this, we inject reference appearance through region-restricted semantic feature blending in a pretrained DiffAE encoder. To ensure geometric compatibility, we further apply landmark-driven thin-plate spline (TPS) warping to adapt the reference part to the target's per-frame facial configuration. Finally, stochastic boundary blending restores fine details, and optical-flow-guided regional attention enforces temporal consistency.

To enable systematic evaluation, we construct a benchmark based on FaceForensics++ [16] with 100 target identi-

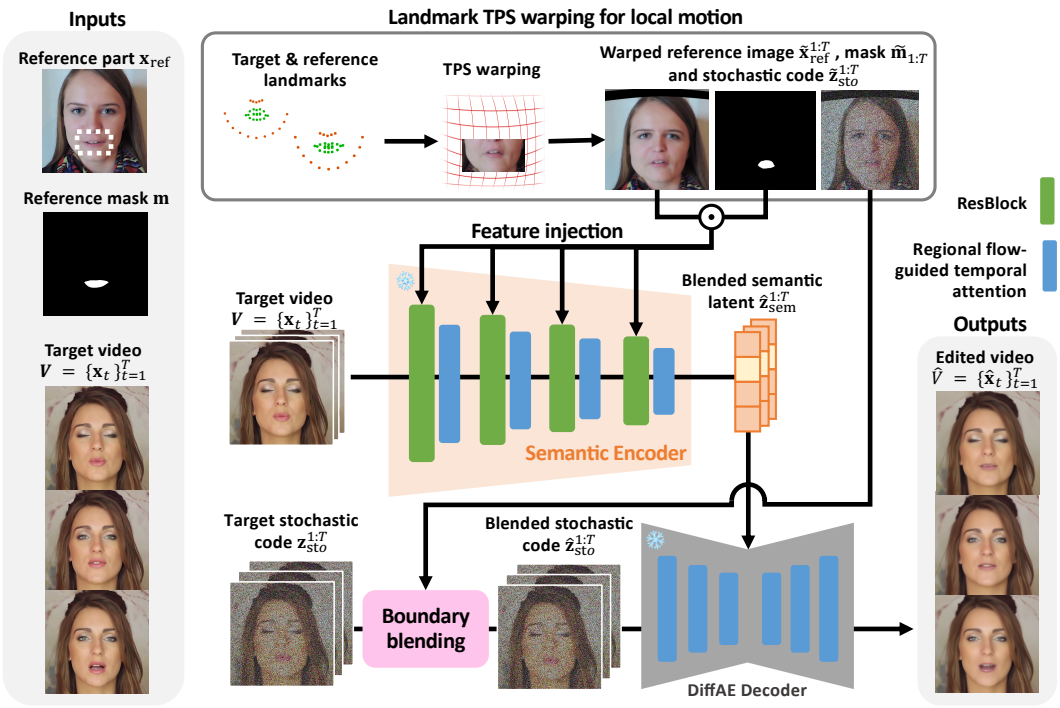

Figure 2. **Overview of V-PartSwap.** Our pipeline combines (i) semantic feature blending, (ii) landmark-guided TPS alignment, (iii) stochastic boundary blending, and (iv) regional flow-guided temporal attention for motion-consistent facial part transfer.

ties and reference parts for eyebrows, eyes, nose, and mouth. Experiments show that our approach produces motion-consistent edits while preserving overall facial realism.

2. Related Works

Facial part transfer aims to replace a specific facial component of a target face using a reference image while preserving the remaining facial regions. Existing methods mainly address this task in static images using GAN or diffusion-based frameworks, including region-based style transfer, latent editing, and part-aware fusion [7, 13, 20, 21, 23]. However, these approaches operate on individual frames and do not explicitly model temporal consistency.

Full-face swapping has been extensively studied in both images and videos. Early works rely on GAN-based identity injection [2, 14], while recent diffusion-based methods improve realism and controllability through conditional generation [9, 12]. Video extensions further address identity–motion disentanglement and temporal coherence [1, 8, 18]. In contrast, facial part transfer requires localized identity injection while preserving global identity and motion.

Our method builds on Diffusion Autoencoders (DiffAE) [15], which enable controllable editing through a structured semantic–stochastic latent space.

3. Method: V-PartSwap

Given a target video $V = \{\mathbf{x}_t\}_{t=1}^{T}$, a reference image $\mathbf{x}_{\text{ref}}$, and a reference mask $\mathbf{m}$, our goal is to synthesize an edited video $\hat{V} = \{\hat{\mathbf{x}}_t\}_{t=1}^{T}$ in which the selected facial part matches the reference appearance while preserving the target identity and motion.

As shown in Fig. 2, our pipeline consists of four steps: (i) region-aware semantic feature blending in DiffAE, (ii) landmark-guided TPS warping to align the reference part with the target motion, (iii) boundary-aware blending of the stochastic latent code for detail recovery, and (iv) regional flow-guided temporal attention for temporal consistency.

3.1. Semantic Feature Injection

Diffusion autoencoder. We build on DiffAE [15], which decomposes an image into a semantic latent $\mathbf{z}_{\text{sem}}$ and a stochastic latent $\mathbf{z}_{\text{sto}}$. The semantic code captures high-level facial structure, while the stochastic code preserves residual details. For each frame $\mathbf{x}_t$, we obtain $\mathbf{z}_{\text{sem}}^{t} = \text{Enc}(\mathbf{x}_t)$, $\mathbf{z}_{\text{sto}}^{t} = \text{DDIM\_inv}(\mathbf{x}_t, \mathbf{z}_{\text{sem}}^{t})$.

Region-aware semantic feature blending. To transfer a specific facial part while preserving the remaining regions, we inject reference semantics only inside the target part mask. Let $\mathbf{h}_t^i$ and $\mathbf{h}_{\text{ref}}^i$ denote intermediate encoder features of the target and reference at resolution i. Given a resized mask $\mathbf{m}^i$, we perform residual blending

$$\tilde{\mathbf{h}}_t^i = \mathbf{h}_t^i + \mathbf{m}^i \odot (\mathbf{h}_{\text{ref}}^i - \mathbf{h}_t^i), \tag{1}$$

which injects reference semantics only within the selected region. Applying this operation across multiple resolutions yields a semantic latent $\hat{\mathbf{z}}_{\text{sem}}$ encoding the reference part identity.

Figure 3. Qualitative comparison of facial-part transfer in videos. Each example shows the target video, reference image, and edited outputs compared with our baselines.

3.2. Landmark TPS Warping for Local Motion

Facial parts such as the mouth exhibit non-rigid motion that cannot be captured by global head pose alone. To align the reference appearance with the target motion, we warp the reference image, mask, and stochastic latent code to each frame using landmark-guided thin-plate spline (TPS) warping. Let $\{\mathbf{p}_j\}_{j\in\mathcal{I}}$ and $\{\mathbf{q}_j^t\}_{j\in\mathcal{I}}$ denote the reference and target landmark coordinates for a selected facial region $\mathcal{I}$ at frame t. We estimate a TPS mapping $\Phi_t(\mathbf{u})$ that aligns $\mathbf{q}_j^t$ to $\mathbf{p}_j$, producing a sampling grid $\mathbf{G}_t$. We warp the reference image $\mathbf{x}_{\text{ref}}$, mask $\mathbf{m}$, and stochastic latent $\mathbf{z}_{\text{sto}}^{\text{ref}}$ by the grid $\tilde{\mathbf{x}}_{\text{ref}}^t, \tilde{\mathbf{m}}_t, \mathbf{z}_{\text{sto}}^{\text{ref}\to t} = \mathcal{W}(\cdot, \mathbf{G}_t)$. This alignment transfers the target's local deformation to the reference part, enabling motion-consistent editing.

3.3. Boundary Stochastic Blending

Replacing the stochastic latent code inside the entire mask can introduce artifacts. Instead, we blend stochastic latents only within a narrow band around the warped mask boundary. Given the TPS-warped mask, we construct a soft boundary weight map $\tilde{\mathbf{b}}_t$ using morphological operations followed by Gaussian smoothing. Let $\mathbf{z}_{\text{sto},t}^{\text{tgt}}$ denote the target stochastic latent and $\mathbf{z}_{\text{sto}}^{\text{ref}\to t}$ the TPS-warped reference latent. We compute the blended latent as $\hat{\mathbf{z}}_{\text{sto}}^t = (1-\tilde{\mathbf{b}}_t)\odot \mathbf{z}_{\text{sto},t}^{\text{tgt}} + \tilde{\mathbf{b}}_t \odot \mathbf{z}_{\text{sto}}^{\text{ref}\to t}$, allowing fine details to be restored while maintaining smooth transitions.

3.4. Regional Flow-guided Temporal Attention

To improve temporal stability, we apply a flow-guided temporal attention module following FLATTEN [3]. The attention is restricted to a region of interest defined by the warped mask, allowing features to be aggregated only from motion-corresponding locations across frames without affecting unedited regions.

4. Experiments

We describe our evaluation benchmark and then present the experiments.

4.1. Experiment Setup

Evaluation benchmark. Since no benchmark exists for video facial part transfer, we construct one based on FaceForensics++ [16]. We select clips containing at least 120 consecutive frames and randomly sample 100 videos as targets. For reference images, we curate a pool for four facial parts (eyebrows, eyes, nose, and mouth). Candidate frames are sampled from videos that do not overlap with the target identities. Each frame undergoes face alignment and semantic parsing [24] to extract the corresponding part mask. Quality filters are applied to remove blurred images, truncated masks, and extreme expressions. The final reference pool contains 320 aligned samples (80 per part), each consisting of an RGB image, part mask, and facial landmarks.

Table 1. Quantitative comparison of the common subset of facial parts (Eye, Mouth, Nose), which are supported by all evaluated methods. ROI-based fidelity metrics (CLIP-I, DINO-I) are macro-averaged over the three parts. Identity and background metrics are computed outside the edited ROI. Temporal and global consistency metrics are computed on the full edited videos corresponding to the same subset. LPIPS_{BG} denotes the background LPIPS. E_{warp} is reported in 10^{-3}. ↑ indicates higher is better and ↓ indicates lower is better. Best results are **bold**, and second-best results are underlined.

Method	Quality		Motion		Editing Fidelity		Identity	
	E_{warp} ↓	FVD ↓	Pose ↓	Expr. ↓	CLIP-I ↑	DINO-I ↑	ArcFace ↑	LPIPS_{BG} ↓
FacePartsSwap [5]	9.232	30.723	1.756	0.142	**0.579**	**0.384**	0.929	0.160
E4S [13]	9.439	39.249	1.802	0.141	0.568	0.375	0.632	0.215
DiffSwap [23]	9.452	60.043	1.814	0.116	0.563	0.364	0.944	0.0320
FuseAnyPart [20]	9.618	72.524	2.032	0.188	0.563	0.359	0.770	0.0737
VideoSwap [6]	9.351	212.41	2.215	0.195	0.554	0.344	0.482	0.158
VACE [11]	9.232	42.935	1.728	**0.0762**	0.561	0.357	**0.973**	**0.0261**
HunyuanCustom [10]	9.244	**23.379**	**1.727**	0.145	0.574	0.366	0.959	0.0441
MoCha [22]	9.626	119.97	2.965	0.289	0.577	0.365	0.0804	0.341
Ours	**9.143**	27.848	1.919	0.135	0.565	0.371	0.868	0.0693

Implementation Details. We implement our method on the DiffAE FFHQ256 model [15]. All videos and reference images are aligned using Dlib and resized to 256^2. Optical flow is estimated using RAFT [19]. TPS warping is disabled for nose transfer, as nasal variation is mainly governed by global pose and exhibits minimal non-rigid deformation. All experiments are conducted on NVIDIA A100 GPUs. Editing a 120-frame video takes approximately 305 seconds. Our method operates in a fully training-free manner.

Baselines. We compare with representative facial part swapping methods, including FacePartsSwap [5], E4S [13], DiffSwap [23], and FuseAnyPart [20]. We also include general reference-based video editing systems, including VideoSwap [6], VACE [11], HunyuanCustom [10], and MoCha [22]. We use the official implementations and pre-trained models for all baselines. We do not include the results of eyebrow transfer for FuseAnyPart [20], as it requires retraining of the method.

Evaluation Metrics. We evaluate editing quality along four aspects. *Temporal and video quality:* Optical-flow warping error (E_{warp}) and Fréchet Video Distance (FVD). *Motion preservation:* Pose error (HopeNet [17]), and expression error (Deep3DFaceRecon [4]). *Editing fidelity:* CLIP-I and DINO-I similarity between edited parts and reference parts. *Identity preservation:* ArcFace similarity and background LPIPS computed outside the edited region.

4.2. Main Results

Visual comparisons. Figure 3 presents qualitative comparisons with representative baselines. DiffSwap often fails to fully replace the target attributes, while FacePartsSwap frequently produces visible boundary artifacts. VACE tends to generate weak edits for fine-grained attributes, such as eyebrows and the nose, whereas HunyuanCustom sometimes produces appearance mismatches due to differences in skin tone or illumination. In contrast, our method consistently transfers the desired facial parts while maintaining natural appearance and boundary coherence.

Quantitative comparisons. Table 1 reports quantitative results on the common subset of facial parts (eyes, mouth, and nose). Our method achieves the lowest warping error E_{warp}, indicating stronger motion consistency between the edited regions and the target video. It also obtains the second-best FVD score, suggesting improved perceptual video quality. Despite being training-free, our approach maintains appearance fidelity (CLIP-I and DINO-I) comparable to recent methods, while preserving identity and background consistency across frames.

Image-based part transfer methods such as FacePartsSwap and E4S achieve strong appearance fidelity but often introduce boundary artifacts. Video editing approaches such as VideoSwap and MoCha exhibit larger identity drift, reflected by lower ArcFace similarity. Diffusion-based part-transfer methods achieve competitive fidelity but exhibit poorer overall quality. Overall, our method achieves stronger motion alignment while maintaining competitive appearance fidelity and identity preservation.

5. Conclusions

This paper introduced V-PartSwap, a training-free framework for motion-consistent facial part transfer in videos. Built upon DiffAE, our method performs localized semantic blending in the latent space while preserving non-edited facial regions. By incorporating landmark-guided alignment and region-guided temporal attention, V-PartSwap preserves reference appearance while maintaining compatibility with the target face and temporal consistency across frames.

The method does not explicitly model teeth, which may lead to artifacts when the mouth opens widely, and large appearance mismatches between reference and target identities may produce overly smoothed results. Addressing these challenges is left for future work.

References

[1] Sanoojan Baliah, Yohan Abeysinghe, Rusiru Thushara, Khan Muhammad, Abhinav Dhall, Karthik Nandakumar, and Muhammad Haris Khan. Vface: A training-free approach for diffusion-based video face swapping. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4315–4324, 2026. 2

[2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACM MM*, pages 2003–2011, 2020. 2

[3] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical flow-guided attention for consistent text-to-video editing. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 3

[4] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 4

[5] Claudio Ferrari, Matteo Serpentoni, Stefano Berretti, and Alberto Del Bimbo. What makes you, you? analyzing recognition by swapping face parts. In *Proceedings of IEEE International Conference on Pattern Recognition (ICPR)*, pages 945–951. IEEE, 2022. 4

[6] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7621–7630, 2024. 4

[7] Jingtao Guo and Yi Liu. Facial parts swapping with generative adversarial networks. *Journal of Visual Communication and Image Representation*, 78:103152, 2021. 2

[8] Xu Guo, Fulong Ye, Xinghui Li, Pengqi Tu, Pengze Zhang, Qichao Sun, Songtao Zhao, Xiangwang Hou, and Qian He. Dreamid-v: Bridging the image-to-video gap for high-fidelity face swapping via diffusion transformer. *arXiv preprint arXiv:2601.01425*, 2026. 2

[9] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2024. 2

[10] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 4

[11] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[12] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *PR*, 163:111451, 2025. 2

[13] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8578–8587, 2023. 2, 4

[14] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7184–7193, 2019. 2

[15] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10619–10629, 2022. 1, 2, 4

[16] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1, 3

[17] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, 2018. 4

[18] Hao Shao, Shulun Wang, Yang Zhou, Guanglu Song, Dailan He, Zhuofan Zong, Shuo Qin, Yu Liu, and Hongsheng Li. Vividface: A robost and high-fidelity video face swapping framework. In *NeurIPS*, 2025. 2

[19] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 4

[20] Yaohua Wang, Siying Cui, Aixi Zhang, Wei-Long Zheng, Senzhang Wang, et al. Fuseanypart: Diffusion-driven facial parts swapping via multiple reference images. In *NeurIPS*, 2024. 1, 2, 4

[21] Zhiliang Xu, Xiyu Yu, Zhibin Hong, Zhen Zhu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Facecontroller: Controllable attribute editing for face in the wild. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 3083–3091, 2021. 2

[22] Zhengbo Xu, Jie Ma, Ziheng Wang, Zhan Peng, Jun Liang, and Jing Li. End-to-end video character replacement without structural guidance. *arXiv preprint arXiv:2601.08587*, 2026. 4

[23] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8568–8577, 2023. 2, 4

[24] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18697–18709, 2022. 3