

FlashLips: 100-FPS Mask-Free Latent Lip-Sync using Reconstruction Instead of Diffusion or GANs

Supplementary Material

A. Training Details

A.1. Data Augmentation

During training for both stages, we apply the following augmentations. All images are normalized by dividing by 255 to map pixel values into the range $[0, 1]$. For Stage 1, we additionally apply random horizontal flips with probability 0.5 and ColorJitter with a coefficient of 0.4 for hue, brightness, and contrast, 3.2 for saturation, and an overall application probability of 0.9. For the distilled mouth-latent network, we downscale the source image to 384×384 using interpolation.

A.2. Mask Removal Training Details

After the reconstruction editor R_ϕ converges, we synthesize lip-altered counterparts for real frames. Given a real frame S and a sampled lips-pose vector \mathbf{z}_{lips} , we produce $\tilde{S} = R_\phi(S; \mathbf{z}_{\text{lips}})$ and form symmetric pseudo-pairs $(S \rightarrow \tilde{S})$ and $(\tilde{S} \rightarrow S)$. We initialize the *LipsChange* editor $L_\theta \leftarrow R_\phi$ and fine-tune it on these pairs using the same objective as in Section 3.1.3. At test time, L_θ runs *without* any explicit mouth masks.

Why two directions?

- $(S \rightarrow \tilde{S})$ (**real** \rightarrow **synth**): input matches inference (real frames), which preserves lip-audio sync; however, the target \tilde{S} may contain minor artifacts from R_ϕ , so training *only* on this direction can reproduce them.
- $(\tilde{S} \rightarrow S)$ (**synth** \rightarrow **real**): target is clean (real S), which discourages artifacts; but the input is synthetic and does not match inference, so training *only* on this direction hurts sync/generalization.

Mixture. We tried various training strategies, but eventually we train our model on a mixture of both directions: sample $(S \rightarrow \tilde{S})$ with probability 2/3 and $(\tilde{S} \rightarrow S)$ with probability 1/3. Empirically, this preserves lip-audio alignment (same LipScore as the real \rightarrow synth-only variant) while improving visual quality by avoiding propagation of reconstruction artifacts. This self-refinement removes the need for external segmentation at inference and keeps the pipeline mask-free.

A.3. Architecture Details

All Stage 1 editors (Reconstruction and *LipsChange*) operate on the SDXL VAE latent grid (stride 8). The *input*

is the channel-wise concatenation of the masked target latent, the identity-adapted reference latent $f_{\text{ref}}(\mathbf{z}_{\text{ref}})$, and the lips-pose vector tiled to the latent resolution; this totals **52** channels in our implementation. The network predicts a **4-channel latent residual** that is added to the masked latent and decoded by the frozen VAE. *LipsChange* shares the same backbone and is initialized from the Reconstruction network weights for mask-free self-refinement.

U-Net. Table A.1 summarizes the U-Net model: we use ResNet2D blocks (GroupNorm (GN) 32, SiLU, 3×3 convolutions). The down path increases the number of channels from $384 \rightarrow 512 \rightarrow 640$ using average-pooling for downsampling; the up path mirrors this structure with skip concatenations and resize-conv upsampling, ending with GN+SiLU+Conv to 4 channels. This backbone yields the best throughput (see Table 2) while preserving identity and background consistency and confining edits to the mouth.

Stage	Composition	Channels (in \rightarrow out)
Input	Conv2d($k=3, s=1, p=1$)	52 \rightarrow 384
Down Block 1	$4 \times \text{ResNet2D} + 1 \times \text{ResNet2D}$ (with AvgPool Downsample)	384 \rightarrow 384
Down Block 2	$4 \times \text{ResNet2D} + 1 \times \text{ResNet2D}$ (with AvgPool Downsample)	384 \rightarrow 512
Down Block 3	$4 \times \text{ResNet2D}$ (no downsampling)	512 \rightarrow 640
Mid Block	$2 \times \text{ResNet2D}$	640 \rightarrow 640
Up Block 1	$5 \times \text{ResNet2D} + 1 \times \text{ResNet2D}$ (with Upsample)	1280 \rightarrow 640
Up Block 2	$5 \times \text{ResNet2D} + 1 \times \text{ResNet2D}$ (with Upsample)	1152 \rightarrow 512
Up Block 3	$5 \times \text{ResNet2D}$ (no upsampling)	896 \rightarrow 384
Output	GN(32) + SiLU + Conv2d($k=3, s=1, p=1$)	384 \rightarrow 4

Table A.1. **Architecture of the U-Net Base Model.** Each ResNet2D block consists of GroupNorm (GN, 32 groups), SiLU activation, and two Conv2d layers ($k=3, s=1, p=1$).

Transformer. Table A.2 summarizes the ViT-style model: a 1×1 input projection ($52 \rightarrow 128$), followed by GN and a 1×1 lift to 1024 channels; then 16 BasicTransformerBlocks (LayerNorm, MHSA with 16 heads \times 64 dim, GEGLU MLP with $4 \times$ expansion); followed by 1×1 projections back to 4 output channels. Convolutional pre-/post-projections preserve the 2D grid, while attention improves global consistency at the cost of lower FPS.

Stage	Composition	Channels (in \rightarrow out)
Input Projection	Conv2d($k=1, s=1$)	52 \rightarrow 128
Transformer Pre-projection	GroupNorm(32) + Conv2d($k=1, s=1$)	128 \rightarrow 1024
Transformer Blocks	$16 \times \text{BasicTransformerBlock}$: LayerNorm + MHSA (16 heads, 64-dim/head) + LayerNorm + FeedForward (GEGLU, $4 \times$ expansion)	1024 \rightarrow 1024
Transformer Post-projection	Conv2d($k=1, s=1$)	1024 \rightarrow 128
Output Projection	Conv2d($k=1, s=1$)	128 \rightarrow 4

Table A.2. **Architecture of the Transformer Base Model.** MHSA stands for Multi-Head Self-Attention.

Trade-off. Both backbones achieve comparable accuracy (main paper). The transformer is slightly stronger on perceptual metrics, whereas the U-Net is substantially faster. This makes the U-Net preferable for real-time use and the transformer preferable for peak visual quality. The Stage-2 flow-matching transformer (FMT) architecture – shared by both U-Net and Transformer variants of FlashLips – is detailed in Table A.3.

Stage	Composition	Dims (in \rightarrow out)
Input Motion Embedding	SequenceEmbed: Linear($d_p \rightarrow d_h$) (no affine norm)	$d_p \rightarrow d_h$ 12 \rightarrow 1024
Positional Encoding	Fixed sinusoidal encoding (non-learnable), added to token embeddings ($T = 60$ frames)	$T \times d_h \rightarrow T \times d_h$ $60 \times 1024 \rightarrow 60 \times 1024$
Time Embedding	TimestepEmbedder: sinusoidal (256-dim) + MLP: Linear($256 \rightarrow d_h$) + SiLU + Linear($d_h \rightarrow d_h$)	256 $\rightarrow d_h$ 256 \rightarrow 1024 1024 \rightarrow 1024
Condition Embedding	Concat of identity, audio and emotion latents: $[w_r, w_a, w_e]$ with Linear($d_{\text{cond}} \rightarrow d_h$)	($d_{\text{cond}} \rightarrow d_h$) $d_{\text{cond}} = d_p \cdot n_{\text{id}} + d_a + d_e$ ($12n_{\text{id}} + 512 + 7$) \rightarrow 1024
FMT Blocks	8 \times FMTBlock: AdaLN-modulated MHSA (8 heads, 128-dim/head) + AdaLN-modulated MLP MLP: Linear($d_h \rightarrow 4d_h$) + GELU + Linear($4d_h \rightarrow d_h$) AdaLN MLP: SiLU + Linear($d_h \rightarrow 6d_h$)	$d_h \rightarrow d_h$ 1024 \rightarrow 1024 1024 \rightarrow 4096 4096 \rightarrow 1024 1024 \rightarrow 6144
Output Decoder	AdaLN: LayerNorm (no affine) + SiLU + Linear($d_h \rightarrow 2d_h$) Linear($d_h \rightarrow d_w$)	$d_h \rightarrow d_h$ 1024 \rightarrow 2048 1024 \rightarrow 12

Table A.3. **Architecture of the Flow Matching Transformer (FMT).** d_p is the lips-pose latent dimension, d_h is the hidden size. MHSA stands for Multi-Head Self-Attention.

B. User Study

To complement our quantitative evaluation, we conducted a user study comparing FlashLips with several baseline lip-sync models. Using the same 100 cross-audio videos as in our quantitative experiments, we present participants with two videos per trial: one generated by our method and one by a randomly selected baseline. Users evaluate either *Visual Quality* or *Lip Sync*, choosing the preferred video or indicating that both are of the same quality. We collect up to 700 votes per baseline comparison and setting, which are then aggregated. The results are shown in Figure C.1.

Across nearly all baselines, FlashLips is the clear user preference for both visual quality and lip-sync accuracy, with a substantial portion of responses also indicating comparable quality. We outperform DiffDub, Diff2Lip, TalkLip, and IP-LAP by a large margin, with only a minority of votes favoring the competing models. Against LatentSync, most users judge the outputs to be similar, with a slight preference for our method. KeySync – a considerably slower (by $\times 30.4$ times, see Table 2) iterative diffusion model – shows a negligible advantage with 29.0% vs 32.7% of votes for visual quality and 26.6% vs 28.4% for lip-sync, although the vast majority of users still deem the two videos to be of equal quality with 38.3% and 45.0% of votes in the respective settings. We attribute this small disadvantage to

artifacts introduced by the SDXL VAE under certain framings and head poses (see Section D).

Overall, the study highlights that FlashLips delivers competitive or superior perceptual quality while operating orders of magnitude faster than state-of-the-art diffusion-based approaches.

C. Additional Quantitative Results

C.1. Mask Removal: Quantitative Impact

To isolate the effect of removing explicit mouth masks, Table C.5 compares the *Transformer with Mask* to our mask-free editors *Transformer Mask-free* and *U-Net Mask-free* under identical evaluation protocols. We treat reconstruction as a sanity check and focus primarily on cross-audio, which reflects the real use case of our model.

Reconstruction. Removing the mask improves both fidelity and lip-sync quality for the Transformer variant: LipScore increases from roughly 0.50 to 0.70–0.75, and all fidelity metrics (FID/FVD, LPIPS, PSNR, SSIM, ID) move in the expected direction, suggesting better distribution match, sharper frames and higher identity similarity. This confirms that mask-free editing can localize mouth modifications without sacrificing reconstruction quality.

Cross-audio. Mask removal yields the largest improvements in cross-audio. For the Transformer, FID drops from ~ 10.2 to ~ 5.7 and FVD from ~ 74 to ~ 25 –41, indicating cleaner frames and substantially more stable motion. ID improves from ~ 0.77 –0.79 to ~ 0.81 –0.82, and HyperIQA increases slightly. LipScore remains in the same range of 0.35–0.40, showing that lip–audio alignment is preserved. Qualitatively, mask-free models reduce mouth glitches and flicker while providing more stable backgrounds and facial detail. The mask-free U-Net follows the same trend, with slightly worse FID/FVD but higher throughput.

Takeaway. Mask-free self-refinement is a key contributor to the final system: it removes the need for segmentation at inference and consistently improves perceptual quality, temporal smoothness, and identity preservation, while maintaining lip–audio alignment comparable to or better than the masked baseline.

C.2. VBench Results

Table C.4 summarizes VBench scores (see Section 4.3). Across both reconstruction and cross-audio settings, our mask-free models achieve the highest or near-highest total score, demonstrating strong subject and background consistency, motion smoothness, and perceived visual fidelity.

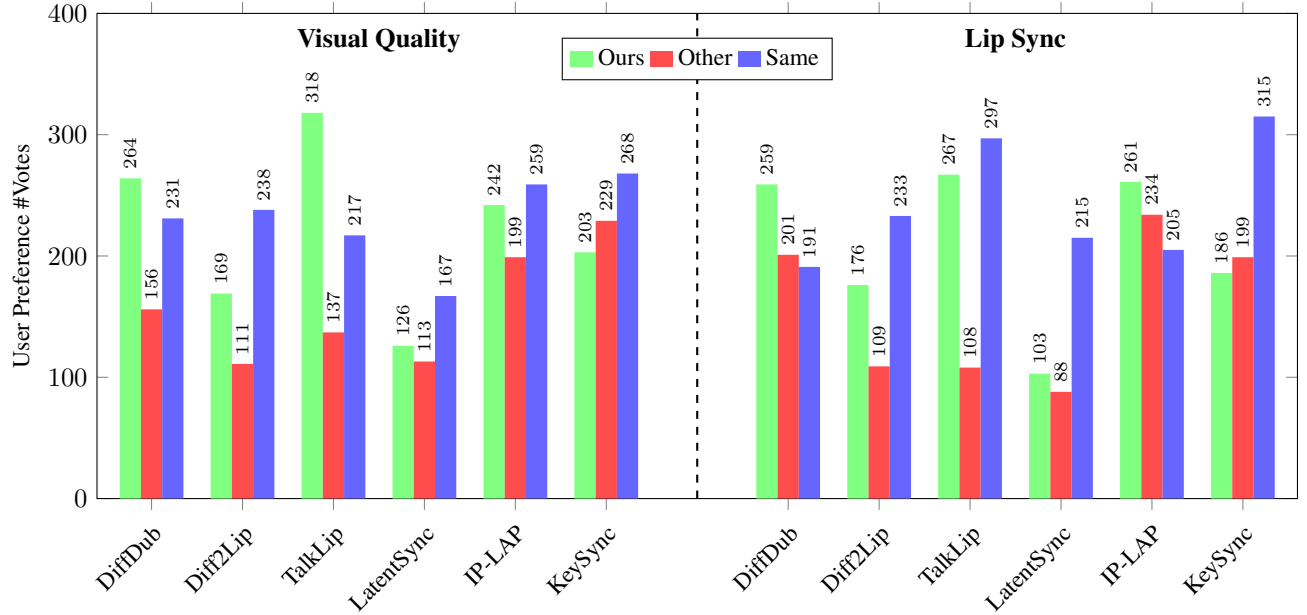


Figure C.1. **Human Preference Evaluation.** We conducted a user study comparing FlashLips against randomly selected baseline models. Participants indicated their preference across two criteria: Visual Quality and Lip Sync. The chart displays the number of responses favoring our model (Ours), the competing model (Other), or neither (Same).

Model	Reconstruction						
	SC \uparrow	BC \uparrow	MS \uparrow	DD \uparrow	AQ \uparrow	IQ \uparrow	Total \uparrow
DiffDub [28]	0.962	0.954	0.992	0.670	0.505	0.661	0.673
Diff2Lip [32]	0.953	0.946	0.992	0.653	0.557	0.633	0.672
TalkLip [49]	0.952	0.942	0.992	<u>0.727</u>	0.528	0.596	0.667
LatentSync [26]	0.967	0.948	<u>0.991</u>	0.723	0.528	0.671	<u>0.682</u>
IP-LAP [64]	0.961	0.945	0.992	0.720	0.519	0.640	0.674
KeySync [2]	0.953	0.948	<u>0.991</u>	0.750	0.531	<u>0.669</u>	0.681
FlashLips – U-Net (Ours)	0.957	0.956	0.990	0.750	<u>0.559</u>	0.667	0.687
FlashLips – Transformer (Ours)	0.957	<u>0.955</u>	0.990	0.750	0.560	<u>0.669</u>	0.687

Model	Cross-Audio						
	SC \uparrow	BC \uparrow	MS \uparrow	DD \uparrow	AQ \uparrow	IQ \uparrow	Total \uparrow
DiffDub [28]	0.956	0.953	0.992	0.624	0.506	0.660	0.668
Diff2Lip [32]	0.946	0.945	<u>0.991</u>	0.622	0.550	0.631	0.667
TalkLip [49]	<u>0.958</u>	0.947	0.992	0.420	0.527	0.596	0.645
LatentSync [26]	0.963	0.953	<u>0.991</u>	0.690	0.537	0.664	0.680
IP-LAP [64]	0.963	0.945	0.992	0.670	0.519	0.639	0.670
KeySync [2]	0.951	0.949	<u>0.991</u>	<u>0.680</u>	0.529	0.668	0.676
FlashLips – U-Net (Ours)	0.955	0.958	0.990	0.670	<u>0.558</u>	<u>0.666</u>	<u>0.681</u>
FlashLips – Transformer (Ours)	0.955	<u>0.957</u>	0.990	<u>0.680</u>	0.559	0.668	0.682

Table C.4. **Quantitative Comparison on VBench.** Video quality evaluation using VBench [17] metrics on 100 randomly sampled reconstruction videos and 100 cross-audio pairs from HDTF, CelebV-HQ and CelebV-Text. Metrics defined in Section 4.3.

D. Limitations

Although our model produces high-quality lip-sync in most cases, it still exhibits some limitations (Figure C.2). Since the method relies on direct prediction rather than the iterative denoising used in diffusion-based approaches, it can struggle to generate fine-grained facial details, particularly

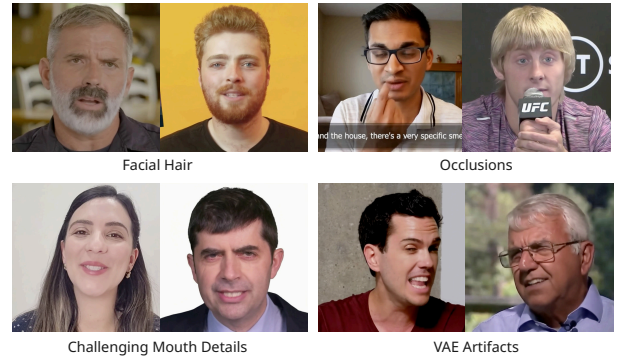


Figure C.2. **Limitations.** Examples illustrating typical failure cases under challenging conditions, including generating facial hair and teeth details, occlusions, and artifacts caused by the SDXL VAE.

in regions such as facial hair and teeth. While the model was not explicitly trained to handle occlusions, it is often surprisingly robust; however, occluding objects can still degrade lip-sync accuracy in more challenging sequences. A more fundamental limitation stems from the SDXL VAE, whose performance degrades in a predictable manner under certain framings and head poses. The VAE performs well on tight close-ups, but when the subject appears in wider shots, artifacts become more common and can adversely affect the lip-sync quality.

Model	# Ref Lats	FID ↓	FVD ↓	HyperIQA ↑	LipScore ↑	ID ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Reconstruction									
Transformer with Mask	1	8.06	57.80	73.26	0.50	0.81	27.68	0.90	0.044
	4	8.01	57.93	73.27	0.53	0.82	27.75	0.90	0.043
	8	7.95	58.15	73.28	0.55	0.82	27.77	0.90	0.043
	16	7.97	57.78	73.29	0.56	0.82	27.79	0.90	0.043
	32	7.94	57.63	73.28	0.55	0.82	27.80	0.90	0.043
Transformer Mask-free	1	4.46	12.53	74.05	0.69	0.85	32.71	0.94	0.021
	4	4.43	12.31	74.06	0.71	0.86	32.88	0.94	0.021
	8	4.41	12.47	74.07	0.73	0.85	33.00	0.94	0.021
	16	4.38	11.90	74.08	0.74	0.86	33.02	0.94	0.021
	32	4.36	12.16	74.08	0.75	0.86	33.10	0.94	0.020
U-Net Mask-free	1	4.73	15.68	73.79	0.67	0.85	32.74	0.94	0.022
	4	4.75	15.20	73.81	0.70	0.85	32.86	0.94	0.022
	8	4.76	15.61	73.82	0.70	0.85	32.92	0.94	0.022
	16	4.66	15.07	73.83	0.71	0.85	32.95	0.94	0.022
	32	4.70	15.85	73.83	0.69	0.85	32.97	0.94	0.022
Cross-Audio									
Transformer with Mask	1	10.24	73.91	73.16	0.40	0.77	—	—	—
	4	9.87	68.10	73.23	0.39	0.77	—	—	—
	8	9.74	68.77	73.26	0.38	0.78	—	—	—
	16	9.67	66.17	73.25	0.35	0.79	—	—	—
	32	9.68	64.92	73.25	0.34	0.78	—	—	—
Transformer Mask-free	1	6.25	41.38	73.80	0.40	0.79	—	—	—
	4	5.89	29.40	73.84	0.37	0.81	—	—	—
	8	5.81	29.54	73.88	0.35	0.81	—	—	—
	16	5.73	26.35	73.88	0.34	0.81	—	—	—
	32	5.68	25.16	73.89	0.32	0.82	—	—	—
U-Net Mask-free	1	6.54	42.78	73.51	0.40	0.79	—	—	—
	4	6.23	33.57	73.58	0.38	0.81	—	—	—
	8	6.13	32.64	73.61	0.36	0.80	—	—	—
	16	6.07	31.34	73.63	0.34	0.81	—	—	—
	32	6.13	28.54	73.63	0.32	0.81	—	—	—

Table C.5. **Full Ablation Study.** Ablation study of our mask and mask-free models, and different numbers of references for the audio-to-latent model for reconstruction and cross-audio. Metrics computed on 100 randomly sampled reconstruction videos and 100 cross-audio pairs from HDTF, CelebV-HQ and CelebV-Text.

E. Ethical Considerations and Societal Impact

Lip-sync technology enables applications such as accessibility tools, film and TV dubbing, translation for non-native audiences, expressive avatars, and content creation. It also carries clear risks: malicious users may create deceptive deepfakes, spread misinformation, or impersonate identities. Our method is intended for beneficial use, and we explicitly discourage any harmful or non-consensual deployment. Any system that alters a person’s likeness should obtain explicit, informed consent.

Our model is trained on publicly available datasets that follow their usage guidelines and on an internal dataset collected with participant consent. As with many audio-

visual models, dataset limitations may bring biases across attributes as skin tone, facial structure, language, or accent.

F. Additional Qualitative Results

We provide qualitative reconstruction results in Figure F.3, comparing FlashLips against all baselines. We also assess visual quality across diverse source–driver head-pose combinations, including frontal–frontal, side–side, side–frontal, and frontal–side pairs (Figure F.4). Finally, we show results on out-of-distribution subjects – synthetic human faces and non-human or stylized characters – to demonstrate that our method remains robust and generalizable under these more challenging conditions (Figure F.5).



Figure F.3. **Qualitative Comparison – Reconstruction.** Comparisons with other lip-sync methods for reconstruction. The first row shows the source video; the following rows display the inferred lip-synced videos by each method.

Frontal to Frontal



Side to Side



Side to Frontal



Frontal to Side



Figure F.4. **Lip-sync results across varying facial pose combinations.** Each triplet shows a source video, video corresponding to the audio driver, and the resulting prediction.



Figure F.5. **Lip-sync results on out-of-distribution (OOD) faces.** The top block of images shows results on generated human faces, while the lower block shows results on non-human or stylized faces. Our method maintains consistent lip synchronization and natural articulation across both domains.