WHERE AND WHEN TO FORCE: ROUTED FORCING FOR STREAMING AVATARS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven streaming avatar generation requires real-time synthesis of speech-synchronized videos with dynamic and diverse motion. Self Forcing uses Distribution Matching Distillation (DMD) to distill bidirectional video diffusion models into causal, few-step generators for real-time streaming. However, DMD minimizes a reverse KL divergence, which is inherently mode-seeking: it causes the student to discard high-dynamic modes and collapse onto static outputs, compressing both dynamics and diversity of generated videos. We find that this collapse is *region-heterogeneous*: person regions involving pose and gesture variations suffer the largest diversity loss, the audio-driven mouth region shows a small loss, and the background remains nearly stable. Based on this observation, we propose **Routed Forcing**, which routes the distillation objective by semantic region and noise stage to improve dynamics and diversity while preserving visual quality. Specifically, (1) *Where to Force: Semantic-Region Routing* applies Data-Forcing Distillation (DFD), which supervises the student with real videos, to the person region where diversity collapse is most severe, while retaining DMD for the mouth and background to preserve lip synchronization and scene stability. (2) *When to Force: Noise-Stage Routing* activates DFD at high noise stages, where real video serves as effective supervision to inject diverse and dynamic motion patterns. At low noise stages, DMD is used to refine details, avoiding blur and artifacts from spatial differences between real video and student-generated video. Experiments show that Routed Forcing improves dynamics by up to 45% and diversity by 7–25% over Self Forcing, while preserving video quality and lip synchronization.

1 INTRODUCTION

Audio-driven portrait video generation produces videos whose motion follows input speech (Suwajanakorn et al., 2017; Zhou et al., 2020; Zhang et al., 2023; Tian et al., 2024). Recent methods further extend this paradigm toward causal, real-time streaming generation for interactive applications (Zhen et al., 2025; 2026; Cho et al., 2026; Xiao et al., 2025). This task requires the system to satisfy two simultaneous demands: real-time streaming output at interactive frame rates, and temporally rich dynamics including natural upper-body motion, responsive head turns, and coordinated gestures. Large bidirectional video diffusion models (Yang et al., 2024; Team Wan et al., 2025; Gan et al., 2025; Wang et al., 2025) can already produce high-quality portrait videos, but their multi-step inference cost cannot meet the latency requirements of real-time interaction. Self Forcing (Huang et al., 2025) leverages Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) to distill these models into causal, few-step streaming generators, enabling real-time generation.

However, DMD compresses the dynamics and diversity of generated videos while reducing sampling steps. As shown in Figure 1a, given the same first frame and audio with different random seeds, the original teacher produces varied yet plausible expressions and dynamic gestures, while the DMD-trained student generates highly similar motion trajectories with little hand movement. In other words, the teacher's diverse and dynamic patterns have been compressed into a single, static output. The root cause is that DMD minimizes a reverse KL divergence, which is inherently mode-seeking (Yin et al., 2024b): it penalizes the student for placing mass outside the teacher's support but imposes no penalty for missing teacher modes. *A direct consequence of this diversity collapse is that the student discards high-dynamic modes and falls back to the simplest low-dynamic*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

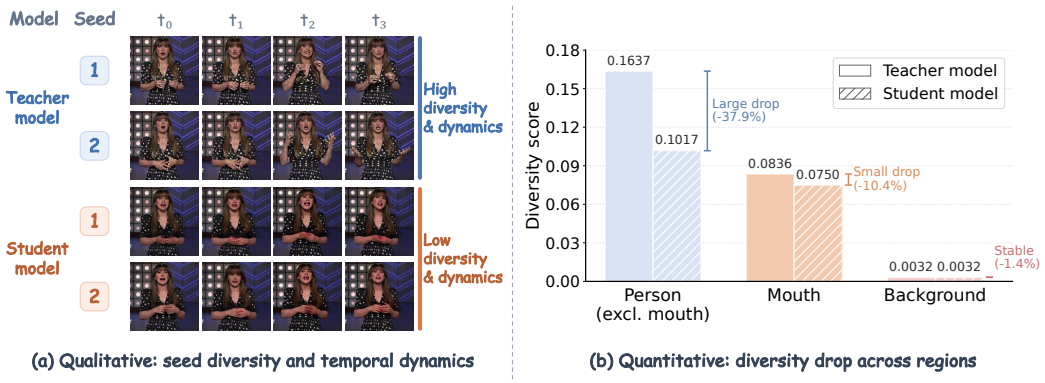


Figure 1: **DMD compresses dynamics and diversity, and the collapse is region-heterogeneous.** We use OmniAvatar as teacher model and LiveTalk, distilled from OmniAvatar via DMD, as student model. (a) Given the same first frame and audio, the teacher produces diverse expressions and dynamic gestures across two seeds, while the student generates similar static trajectories. (b) Regional diversity (mean pairwise LPIPS across four seeds over 100 first-frame-and-audio pairs from SpeakerVid) reveals that the non-mouth person region suffers the largest drop, the mouth region shows a small drop, and the background remains nearly stable.

modes (Huang et al., 2026), resulting in generated outputs that consistently lack natural motion variation, especially harmful for avatars that demand lifelike motion.

Furthermore, we find that this diversity collapse is not uniform across all regions. We measure diversity across the mouth, non-mouth person region, and background over 100 initial conditions, as shown in Figure 1b. The non-mouth person region has many plausible modes for the same speech input, including varied head motions, expressions, and gestures, and suffers the largest diversity loss. The mouth region is strongly constrained by audio, with small diversity loss. The background is nearly deterministic and remains stable. We term this phenomenon *region-heterogeneous collapse*: different semantic regions suffer different degrees of diversity loss under DMD distillation. This regionally heterogeneous structure indicates that addressing DMD-induced diversity collapse should be differentiated according to regional characteristics.

Recently, Data-Forcing Distillation (DFD) (Chen et al., 2026) was proposed to address the diversity collapse caused by DMD. It introduces real-data supervision by querying the real score on condition-matched real video instead of the student rollout video, injecting the high-dynamic, diverse motion patterns from real video into the student. DFD can improve dynamics and diversity, but it applies uniformly across all regions and noise stages, ignoring the regional differences and noise-stage differences. Injecting diversity into regions such as the mouth and background instead disrupts lip accuracy and scene stability. Moreover, at low noise stages, layout differences between real video and the student rollout video (hand position, head angle, body pose) are large, misaligning DFD's gradient with the student's generation and introducing blur and artifacts.

Motivated by the above findings, we propose **Routed Forcing**, which routes DFD along two dimensions to apply it precisely where and when it is needed. **(1) Where to Force: Semantic-Region Routing.** Based on regional diversity differences, Routed Forcing confines DFD to the non-mouth person region where collapse is most severe, while mouth and background regions retain DMD to preserve lip accuracy and scene stability. **(2) When to Force: Noise-Stage Routing.** Routed Forcing further refines DFD application in the non-mouth person region by using DFD and DMD at different noise stages. At high noise stages, real video and the student rollout video are close in noisy space, so DFD is used to inject diverse and dynamic patterns from real video. At low noise stages, DMD is used instead to refine details on the student's own trajectory, avoiding blur and artifacts from cross-trajectory interference.

Experiments show that Routed Forcing improves dynamics by up to 45% and diversity by 7–25% over Self Forcing, while preserving video quality and lip synchronization. Ablation studies confirm the contribution of each routing dimension: semantic-region routing avoids the damage that applying DFD to all regions inflicts on lip accuracy and scene stability, noise-stage routing further eliminates

the blur introduced by layout differences at low noise stages, and their combination yields the best overall performance. Our contributions are summarized as follows:

- We propose Routed Forcing for streaming avatar generation, a simple yet effective training strategy which routes the distillation objective by semantic region and noise stage to improve dynamics and diversity while preserving visual quality.
- We discover *region-heterogeneous collapse* in DMD: diversity loss concentrates in specific semantic regions rather than uniformly across the frame.
- We propose semantic-region routing, which confines DFD to the non-mouth person region, where collapse is most severe, while retaining DMD for the mouth and background to preserve lip accuracy and scene stability.
- We propose noise-stage routing, which activates DFD only at high noise stages to inject diverse and dynamic motion patterns from real video, and uses DMD at low noise stages to refine details and avoid blur and artifacts.

2 Related Work

Audio-driven avatar generation. Early methods animate facial landmarks or 3D morphable models from audio (Suwajanakorn et al., 2017; Chen et al., 2019; Zhou et al., 2020; Thies et al., 2020; Zhang et al., 2023), while recent methods directly generate video frames with diffusion models. AniPortrait (Wei et al., 2024) combines facial landmarks with a diffusion model to produce high-quality portrait animation, FantasyTalking (Wang et al., 2025) further improves expression and motion quality through a diffusion transformer, and OmniAvatar (Gan et al., 2025) introduces adaptive body animation into a bidirectional diffusion transformer for high-fidelity portrait generation. However, these methods require many denoising steps per frame and cannot meet the demands of real-time interaction. LiveTalk (Chern et al., 2025) achieves real-time streaming by distilling a bidirectional teacher into a causal, few-step student via DMD, but the distillation process introduces diversity collapse. *Routed Forcing maintains real-time inference while notably improving dynamics and diversity without sacrificing video quality or lip synchronization.*

Diffusion model distillation. Distilling multi-step diffusion models (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022; Song et al., 2021a; Peebles & Xie, 2023) into few-step generators is key to enabling real-time inference. Progressive distillation (Salimans & Ho, 2022) achieves acceleration by iteratively halving the number of sampling steps, while consistency distillation (Song et al., 2023; Luo et al., 2023) learns a consistency mapping along the PF-ODE trajectory for single- or few-step generation. DMD (Yin et al., 2024b;a) matches the score functions of the student and teacher distributions, evaluated entirely on the student's own samples, producing a mode-seeking (reverse-KL-like) gradient. To mitigate the mode collapse of DMD, DFD (Chen et al., 2026) queries the frozen teacher score on noised real data, so that teacher modes unvisited by the student still provide supervision. *In streaming avatar generation, DMD causes diversity collapse, while global DFD introduces blur at low noise stages. Routed Forcing restricts DFD to the person region at high noise stages, improving diversity where collapse is most severe while avoiding lip-sync degradation.*

Streaming video generation. Self Forcing (Huang et al., 2025) feeds the student's own generated output as the conditioning for subsequent frames and applies DMD on top to align the student and teacher distributions, achieving causal, few-step streaming video generation. Building on this paradigm, several works improve the training strategy. Reward Forcing (Lu et al., 2025) applies exponential reward weighting to the DMD gradient, steering the student distribution toward high-reward samples. DynaForcing (Huang et al., 2026) replaces the rollout starting point with a noised real latent and combines dynamics-aware reward weighting to alleviate dynamic collapse. Causal Forcing (Zhu et al., 2026) further improves distillation quality by bridging the architectural gap between bidirectional teacher and causal student. *These methods focus on improving the rollout, reward signal, or initialization, whereas Routed Forcing explores a different angle by differentiating the distillation objective at the level of semantic regions and noise stages.*

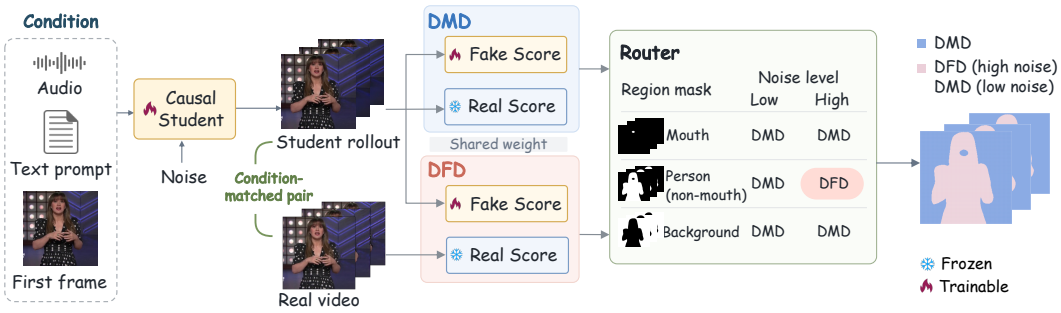


Figure 2: **Routed Forcing routes distillation by semantic region and noise stage.** The causal student generates streaming video from a first frame, audio, and text prompt. DMD queries both the fake score and real score on the student rollout video, while DFD queries the real score on real video with the same condition to inject diverse and dynamic motion patterns from real data. The router selects DFD only for the non-mouth person region at high noise stages and DMD elsewhere.

3 Routed Forcing

We propose Routed Forcing, a simple and effective training strategy that routes the distillation objective by semantic region and noise stage to improve diversity and dynamics while preserving generation quality. We first introduce preliminaries in §3.1, then present the region-heterogeneous collapse phenomenon in §3.2, semantic-region routing in §3.3, and noise-stage routing in §3.4. The overall pipeline is illustrated in Figure 2.

3.1 Preliminaries

Self Forcing. Given a reference image, audio, and text prompt as condition c, the teacher is a bidirectional diffusion transformer that generates high-quality video via multi-step denoising but cannot meet the demands of real-time inference. Self Forcing (Huang et al., 2025) distills it into a causal student that generates frames or chunks autoregressively, enabling streaming generation. During training, the student performs forward inference on its own rollout and aligns its output distribution with the teacher via DMD.

Distribution Matching Distillation (DMD). DMD (Yin et al., 2024b) trains the student by minimizing the reverse KL divergence between the student-induced distribution p_θ and the teacher data distribution p_{data} across noise levels:

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_t \left[D_{\text{KL}}(p_{\theta,t} \| p_{\text{data},t}) \right]. \tag{1}$$

The gradient takes the form of the difference between a fake score and a teacher score, both evaluated on the student's own samples:

$$g_{\text{DMD}}(t) = w_t \left[\hat{x}_{\text{fake}}(z_t^s, c, t) - \hat{x}_T(z_t^s, c, t) \right], \tag{2}$$

where $z_t^s = \alpha_t x^s + \sigma_t \epsilon$ is the noised student rollout, x^s is the clean latent generated by the student, $\hat{x}_T$ is the frozen teacher denoiser, $\hat{x}_{\text{fake}}$ is a learned fake-score model tracking the evolving student distribution, and w_t is a noise-level-dependent weighting coefficient. This gradient is mode-seeking: it penalizes the student for generating outside the teacher's support but provides no signal for teacher modes the student has not visited.

Data-Forcing Distillation (DFD). To mitigate the mode collapse of DMD, DFD (Chen et al., 2026) replaces the teacher's input with a noised real video $z_t^r = \alpha_t x^r + \sigma_t \epsilon$, where x^r is condition-matched real video sharing the same noise ϵ as the student:

$$g_{\text{DFD}}(t) = w_t \left[\hat{x}_{\text{fake}}(z_t^s, c, t) - \hat{x}_T(z_t^r, c, t) \right]. \tag{3}$$

Because x^r is drawn from the full data distribution, teacher modes that the student has never visited still contribute supervision. In practice, DFD starts from a converged DMD checkpoint and switches between DMD and DFD gradients with 50% probability at each training step.

3.2 OBSERVATION: REGION-HETEROGENEOUS COLLAPSE

The mode-seeking nature of DMD leads to diversity loss after distillation, and we find that this loss is not uniform across all regions. We define three semantic regions: mouth, non-mouth person, and background. We quantify diversity for each semantic region by measuring the mean pairwise LPIPS across different random seeds. Results measured with L_1 are provided in Figure 7 in the appendix.

As shown in Figure 1a, the teacher produces diverse gesture variations under the same condition with different seeds, such as raising and waving hands, while the student's two seeds yield nearly identical and static trajectories, with hands showing almost no change and remaining consistent with the first frame. Figure 6 in the appendix further confirms this trend: diversity decreases during DMD training. The quantitative analysis in Figure 1b further reveals the region-heterogeneous nature of this collapse: the non-mouth person region suffers the largest diversity drop of 37.9%, far exceeding the mouth (10.4%), while the background remains nearly stable (1.4%).

We term this phenomenon *region-heterogeneous collapse*: different semantic regions suffer different degrees of diversity loss under DMD distillation. The non-mouth person region has many plausible modes for the same speech input, including varied head motions, expressions, and gestures, and suffers the largest diversity loss. The mouth region is strongly constrained by audio, with small diversity loss. The background is nearly deterministic and remains stable. This regionally heterogeneous structure motivates subsequent region-level differentiation in the distillation objective.

3.3 WHERE TO FORCE: SEMANTIC-REGION ROUTING

Based on the observation in §3.2, Routed Forcing assigns different distillation objectives to different semantic regions according to their collapse severity. Routed Forcing uses masks to assign each spatial location to three semantic regions: mouth (M_m), non-mouth person (M_p), and background (M_b), where $M_m + M_p + M_b = 1$. The three regions have distinct distillation needs. The non-mouth person region (head motion, expression, gestures) suffers the most severe diversity collapse and is where the real-data supervision of DFD is most needed. The mouth region is strongly constrained by audio, with limited diversity collapse, retaining DMD to preserve lip accuracy. The background region is nearly deterministic, with minimal diversity collapse, and similarly uses DMD to maintain scene stability.

Routed Forcing therefore introduces DFD only in the non-mouth person region. Following the original DFD training strategy, this region switches between $g_{\text{DMD}}(t)$ and $g_{\text{DFD}}(t)$ with 50% probability at each step. The expected gradient after region routing is:

$$g = M_p \cdot \tfrac{1}{2}\big[g_{\text{DMD}}(t) + g_{\text{DFD}}(t)\big] + (1 - M_p) \cdot g_{\text{DMD}}(t). \tag{4}$$

However, region routing alone is not sufficient: even when restricted to the non-mouth person region, applying DFD across all noise stages still introduces artifacts at low noise stages. This motivates further routing along the noise-stage dimension.

3.4 WHEN TO FORCE: NOISE-STAGE ROUTING

DFD replaces the real score's input z_t^s with the noised real video z_t^r, where $z_t^s = \alpha_t x^s + \sigma_t \epsilon$, $z_t^r = \alpha_t x^r + \sigma_t \epsilon$, x^s is the clean latent generated by the student, x^r is the condition-matched real video latent, σ_t is the noise coefficient, and both share the same noise ϵ. Since the shared noise cancels, the noise-normalized distance between z_t^s and z_t^r is:

$$\frac{\|z_t^s - z_t^r\|}{\sigma_t} = \sqrt{\text{SNR}_t} \cdot \|x^s - x^r\|, \tag{5}$$

where $\text{SNR}_t = \alpha_t^2/\sigma_t^2$ is the signal-to-noise ratio at timestep t.

At high noise stages (SNR_t small), real video and student rollout video are close in noise space, with small cross-trajectory interference. Denoising at this stage determines coarse-grained semantic trajectories, and introducing real video here injects the rich and dynamic motion patterns into the student, improving the dynamics and diversity. At low noise stages (SNR_t large), the DFD update approaches a direct fit to the paired real trajectory, and cross-trajectory differences (hand positions, head angles, body poses) can blur structural details. Denoising at this stage controls fine details

(texture, hand details, facial identity consistency) and should be guided by DMD on the student's own generation trajectory. We provide a detailed mathematical analysis of why DFD should be restricted to high noise stages in Appendix A.

Based on the above analysis, instead of the 50% random mixing used in original DFD (Eq. 4), we propose Noise-Routed DFD (NR-DFD), which deterministically restricts DFD to high noise stages based on a noise threshold τ:

$$g_{\text{NR-DFD}}(t) = \begin{cases} g_{\text{DFD}}(t) & t \geq \tau \quad \text{(high noise)} \\ g_{\text{DMD}}(t) & t < \tau \quad \text{(low noise)}. \end{cases} \tag{6}$$

Combining region routing and noise routing, the full Routed Forcing gradient is:

$$g_{\text{RF}} = M_p \cdot g_{\text{NR-DFD}}(t) + (1 - M_p) \cdot g_{\text{DMD}}(t). \tag{7}$$

By combining semantic-region routing and noise-stage routing, Routed Forcing precisely applies DFD's real-data supervision where and when it is needed, improving diversity and dynamics while avoiding the degradation in lip accuracy and visual quality caused by global DFD.

4 Experiments

4.1 Experiment Setup

Implementation details. We implement Routed Forcing based on OmniAvatar (Gan et al., 2025). The student is a causal OmniAvatar-1.3B initialized from LiveTalk (Chern et al., 2025) weights. The frozen teacher (real score) is a bidirectional OmniAvatar-14B, and the fake-score model is a bidirectional OmniAvatar-1.3B. We sample 10,000 videos from SpeakerVid (Zhang et al., 2025) as the training set. Following DFD (Chen et al., 2026), we first train 1,000 steps of Self Forcing with DMD to provide a high-quality starting point. Routed Forcing training is then applied for 200 steps. At inference, we generate 512×512 videos of 81 frames at 16 fps, corresponding to 21 latent frames with a causal block size of 3 and four sampling steps per block. Routed Forcing employs Top-50% noise routing, activating DFD only at the top 50% noise timesteps for the non-mouth person region. Person masks are extracted from real videos using DeepLabV3 (Chen et al., 2017), and mouth masks are obtained from student rollouts using InsightFace (Deng et al., 2020). We discuss the mask design choice in Appendix B. More implementation details are provided in Appendix C.

Baselines. We set up two groups of comparisons. (1) *Controlled training strategy comparison*: Self Forcing (Huang et al., 2025) (DMD only), Data-Forcing Distillation (DFD) (Chen et al., 2026) (50% probability DMD, 50% probability DFD at each step), DP-DMD (Wu et al., 2026), DynaForcing (Huang et al., 2026), and Reward Forcing (Lu et al., 2025) (using the reward model from DynaForcing). All methods share the same initialization checkpoint, training data and random seeds, differing only in the algorithm design. (2) *Public model comparison*: We also compare with audio-driven avatar models of comparable scale: LiveTalk (Chern et al., 2025), FantasyTalking (Wang et al., 2025), AniPortrait (Wei et al., 2024), and OmniAvatar (Gan et al., 2025).

Evaluation metrics. We randomly sample 100 videos from SpeakerVid (Zhang et al., 2025), with no overlapping identities with the training set, and 100 from AVSpeech (Ephrat et al., 2018) as test sets. We report ten metrics in four groups. (1) *Video quality*: FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018). (2) *Lip sync*: Sync-C and Sync-D (Chung & Zisserman, 2016). (3) *Dynamics*: RAFT-Motion (Teed & Deng, 2020) (pixel-level motion intensity) and DINO-Temporal (Oquab et al., 2024) (semantic-level temporal variation). (4) *Diversity*: frame-wise mean pairwise distance across four seeds using CLIP (Radford et al., 2021), DINOv2, LPIPS (Zhang et al., 2018), and L1. For comparison with audio-driven avatar models, we additionally report FPS and first-frame latency.

4.2 Main Results

Controlled training strategy comparison. As shown in Table 1, Routed Forcing achieves the best performance across all ten metrics on both datasets. Compared to Self Forcing, Routed Forcing reduces FID by 6.4%/2.5% and FVD by 3.4%/6.9% on SpeakerVid/AVSpeech. The improvements in dynamics are most pronounced: RAFT-Motion increases by 38.8%/45.6% and DINO-Temporal

Table 1: **Controlled comparison of training strategies.** All methods share the same initialization checkpoint, training data and random seeds, differing only in the algorithm design.

Method	Video Quality		Lip Sync		Dynamics (×100)		Diversity (×100)			
	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	RAFT ↑	DINO-T ↑	CLIP ↑	DINO ↑	LPIPS ↑	L1 ↑
SpeakerVid										
Self Forcing	27.17	320.91	4.03	10.97	5.62	1.07	3.07	1.37	4.96	2.14
DFD	28.78	325.33	3.92	11.10	5.47	1.11	3.08	1.42	4.67	1.84
DP-DMD	27.22	320.27	4.00	10.98	5.67	1.06	3.00	1.33	4.89	2.13
Reward Forcing	26.99	315.84	4.04	10.97	5.79	1.09	3.14	1.41	5.20	2.21
DynaForcing	27.75	326.38	4.04	10.95	5.45	1.07	2.96	1.31	4.69	2.06
Routed Forcing	**25.43**	**309.92**	**4.07**	**10.94**	**7.80**	**1.25**	**3.55**	**1.58**	**6.09**	**2.29**
AVSpeech										
Self Forcing	39.44	456.01	3.05	10.77	3.33	0.99	2.42	1.73	4.08	1.90
DFD	41.52	442.15	2.96	10.80	3.46	1.12	2.59	1.89	3.99	1.77
DP-DMD	39.48	454.76	3.02	10.79	3.33	0.98	2.35	1.68	3.97	1.87
Reward Forcing	39.46	453.16	3.04	10.80	3.35	1.01	2.46	1.78	4.17	1.94
DynaForcing	39.78	460.86	3.04	10.77	3.23	0.96	2.29	1.63	3.85	1.82
Routed Forcing	**38.47**	**424.58**	**3.09**	**10.75**	**4.85**	**1.23**	**2.92**	**2.01**	**5.08**	**2.14**



Figure 3: **Qualitative comparison of different training strategies.** All methods are conditioned on the same first frame and audio, and each column shows frames at the same timestep. Routed Forcing produces richer dynamics (e.g., diverse gestures) than Self Forcing and better video quality (e.g., clearer hands) than DFD. More examples are provided in Figure 8 in the appendix.

by 16.8%/24.2%. All four diversity metrics consistently improve by 7–25% on both datasets. Lip sync simultaneously remains the best, indicating that the gains in dynamics and diversity do not come at the cost of lip accuracy.

Moreover, DFD fails to consistently improve video quality (FID is worse than Self Forcing on both datasets) and notably degrades lip sync (Sync-C and Sync-D are both worse than Self Forcing). Globally mixing real-trajectory supervision without distinction hurts both lip accuracy and visual quality, whereas Routed Forcing avoids these side effects through semantic and noise-stage routing. DP-DMD applies teacher-target supervision only at the first generation step, but this per-step routing yields negligible improvement compared to noise-level and region routing. Reward Forcing

Table 2: **Comparison with audio-driven avatar models of comparable scale.**

Method	Size	Efficiency		Video Quality		Lip Sync		Dynamics ($\times 100$)	
		FPS ↑	Latency (s) ↓	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	RAFT ↑	DINO-T ↑
SpeakerVid									
LiveTalk	1.3B	**24.82**	**0.33**	28.77	357.28	3.10	11.62	4.77	1.07
FantasyTalking	14B	0.35	232.06	33.98	364.56	2.44	12.11	3.99	0.57
AniPortrait	2.5B	0.38	211.77	33.49	570.41	3.22	11.24	4.88	1.18
OmniAvatar	1.3B	0.97	83.44	62.83	902.13	1.46	13.74	2.27	**1.70**
Routed Forcing	1.3B	**24.82**	**0.33**	**25.43**	**309.92**	**4.07**	**10.94**	**7.80**	1.25
AVSpeech									
LiveTalk	1.3B	**24.82**	**0.33**	40.00	491.10	2.05	11.79	2.77	1.01
FantasyTalking	14B	0.35	232.06	43.80	461.93	1.74	11.78	3.30	0.66
AniPortrait	2.5B	0.38	211.77	44.76	726.12	1.29	11.88	4.12	**2.03**
OmniAvatar	1.3B	0.97	83.44	76.70	1124.44	0.99	13.13	1.76	2.01
Routed Forcing	1.3B	**24.82**	**0.33**	**38.47**	**424.58**	**3.09**	**10.75**	**4.85**	1.23

and DynaForcing both remain within the mode-seeking DMD framework: Reward Forcing yields only marginal gains through gradient reweighting, and DynaForcing adds data-anchored rollout with reward weighting but still queries scores on student rollouts.

Qualitative comparison of training strategies. As shown in Figure 3, Self Forcing produces near-static avatars with highly consistent gestures, a consequence of DMD mode-seeking that collapses the distribution onto simple static modes. DFD improves dynamics by introducing real-video supervision, but applying DFD at low noise stages causes layout mismatch between student-generated and real trajectories, resulting in noticeable hand blur. Routed Forcing restricts DFD to the non-mouth person region at high noise stages via semantic and noise routing, producing richer dynamics (e.g., diverse gestures) than Self Forcing and better video quality (e.g., clearer hands) than DFD. More examples are provided in Figure 8 in the appendix.

Comparison with avatar models of comparable scale. As shown in Table 2, Routed Forcing shares the same 1.3B causal architecture as LiveTalk, running at 24.82 FPS (0.33 s latency) while improving RAFT-Motion by 63% and FID by 12% on SpeakerVid, showing that training strategy alone yields substantial gains without architectural change. In terms of video quality and lip sync, Routed Forcing notably outperforms all baselines on both datasets. For dynamics, Routed Forcing leads by a large margin in RAFT-Motion (SpeakerVid 7.80 vs. LiveTalk 4.77, FantasyTalking 3.99). OmniAvatar and AniPortrait obtain the highest DINO-Temporal on SpeakerVid (1.70) and AVSpeech (2.03) respectively, but both have lower video quality and lip sync than Routed Forcing.

4.3 Ablation Study

Key component ablation. Table 3 isolates the contributions of region routing and noise routing on SpeakerVid. All four conditions branch from the same SF-1000 checkpoint and are trained for 200 steps. *Region routing* increases RAFT-Motion from 5.62 to 6.27 (+11.6%), while reducing FID from 27.17 to 26.97 and leaving lip sync unchanged, demonstrating that region routing effectively protects mouth accuracy. *Noise routing* yields larger dynamics gains, RAFT-Motion rises from 5.62 to 7.50 (+33.5%) and diversity improves substantially, but Sync-D increases from 10.97 to 11.02, indicating lip-sync degradation when DFD reaches the mouth.

Full Routed Forcing achieves the best FID (25.43) and RAFT-Motion (7.80) among all conditions, while Sync-D not only avoids degradation but reaches the best value. This synergy shows that region routing confines DFD to the non-mouth person region, eliminating the lip-sync side effects of DFD, while noise routing confines DFD to high noise stages, avoiding interference with low-noise details. Combining region routing and noise routing yields improvements across all four metric groups.

Table 3: **Key component ablation.** We isolate the contributions of region routing and noise routing (Top-50%). Region routing applies DFD only to the non-mouth person region. Noise routing activates DFD only at high noise stages.

Self Forcing	Region	Noise	Video Quality		Lip Sync		Dynamics (×100)		Diversity (×100)			
			FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	RAFT ↑	DINO-T ↑	CLIP ↑	DINO ↑	LPIPS ↑	L1 ↑
✓			27.17	320.91	4.03	10.97	5.62	1.07	3.07	1.37	4.96	2.14
✓	✓		26.97	316.81	4.04	10.96	6.27	1.22	3.22	1.51	5.16	2.00
✓		✓	26.61	311.85	4.01	11.02	7.50	1.22	**3.56**	**1.60**	5.92	2.22
✓	✓	✓	**25.43**	**309.92**	**4.07**	**10.94**	**7.80**	**1.25**	3.55	1.58	**6.09**	**2.29**

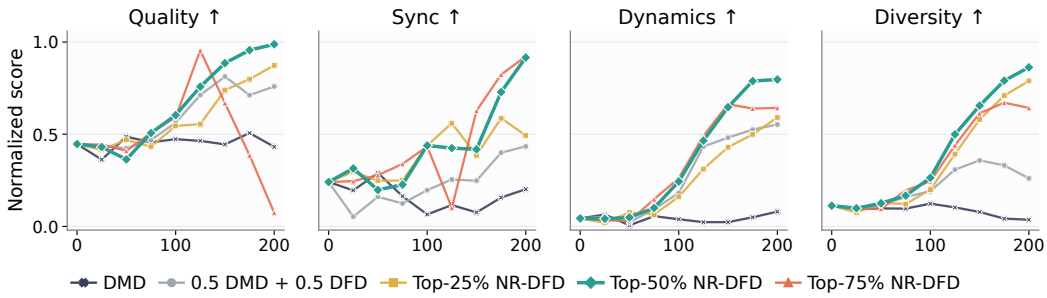


Figure 4: **Noise threshold ablation: training curves on SpeakerVid.** Region routing is kept active throughout: all training variants are applied only to the non-mouth person region, with DMD retained for the mouth and background. The y-axis shows the mean normalized score across metrics.

Noise threshold sensitivity. Figure 4 shows training curves under different noise thresholds. Region routing is kept active throughout: all training variants are applied only to the non-mouth person region, with DMD retained for the mouth and background. Top-k% activates DFD at the highest k% noise stages, and the "0.5 DMD + 0.5 DFD" baseline applies 50% probability DFD mixing across all noise stages, following the original DFD training strategy. The DMD baseline uses no DFD. DMD-only remains stable across all metrics but yields no significant gains. The "0.5 DMD + 0.5 DFD" baseline improves over DMD-only but underperforms noise-routed DFD, confirming that restricting DFD to high noise stages is more effective than uniform mixing. Top-75% exhibits significant quality collapse in later training stages, confirming that DFD at low noise stages harms quality due to layout mismatch between student-generated and real trajectories (consistent with the analysis in §3.4). Top-25% yields improvements between Top-50% and Self Forcing, as the limited DFD proportion restricts its effect. Top-50% strikes the best balance across all four metric groups, outperforming both Self Forcing and the original DFD method.

4.4 Limitations

Semantic-region routing relies on off-the-shelf segmentation models (DeepLabV3 and InsightFace) for mask extraction, so mask accuracy is bounded by the performance of these models. Routed Forcing is validated on streaming avatar generation. The routing principle may generalize to other video generation tasks such as pose-guided human animation, and this remains to be explored.

5 Conclusion

We presented Routed Forcing, which routes DFD by semantic region and noise stage to improve the dynamics and diversity while preserving video quality for streaming avatar generation. Semantic-region routing confines DFD to the non-mouth person region where region-heterogeneous collapse is most severe, retaining DMD elsewhere. Noise-stage routing restricts DFD to high noise stages where cross-trajectory interference is small, with DMD refining details at low noise stages. In the future, we plan to extend Routed Forcing to a broader range of applications beyond streaming avatar generation, such as pose-guided human animation.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript drafting. All AI-generated text was reviewed, revised, and verified by the authors. Generative AI was not used for experimental design, data collection, or result analysis. All experimental results, claims, and conclusions are based on our own implementations and evaluations. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work generates realistic portrait videos from a single reference image and audio, which carries risks of misuse for deepfakes and identity impersonation. We evaluate on publicly available or appropriately licensed datasets and do not collect new human data. We acknowledge the dual-use nature of this technology and encourage responsible deployment with appropriate safeguards such as watermarking and consent verification.

REPRODUCIBILITY STATEMENT

Implementation details are provided in §3 and §4.1, including model architecture, training schedule, hyperparameters, mask extraction pipeline, and evaluation protocol. All controlled comparisons share the same initialization checkpoint, training data, inference settings, and random seeds. We report ten metrics across four groups with exact definitions. Code and trained models will be released upon acceptance.

REFERENCES

Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pp. 7832–7841, 2019.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-Forcing Distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

Ethan Chern, Zhulin Hu, Bohao Tang, Jiadi Su, Steffi Chern, Zhijie Deng, and Pengfei Liu. LiveTalk: Real-time multimodal interactive video diffusion via improved on-policy distillation. *arXiv preprint arXiv:2512.23576*, 2025.

Paul Hyunbin Cho, Jinhyuk Jang, SeokYoung Lee, Joungbin Lee, Siyoon Jin, Heeseong Shin, Jung Yi, Yunjin Park, Chulmin Park, and Seungryong Kim. Lip Forcing: Few-step autoregressive diffusion for real-time lip synchronization. *arXiv preprint arXiv:2606.11180*, 2026.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

Alex Dytso, H. Vincent Poor, and Shlomo Shamai (Shitz). Conditional mean estimation in Gaussian noise: A meta derivative identity with applications. *IEEE Transactions on Information Theory*, 69(3):1883–1898, 2023. doi: 10.1109/TIT.2022.3216012.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4), 2018.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. OmniAvatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025.

Yubo Huang, Sirui Zhao, Xinchen Yao, Zhengye Zhang, Jinyang Huang, Fengqi Cui, Shiwei Wu, and Enhong Chen. DynaForcing: Overcoming dynamic collapse in Self-Forcing distillation for streaming avatar generation. *arXiv preprint arXiv:2608.17707*, 2026.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward Forcing: Efficient streaming video generation with rewarded Distribution Matching Distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4), 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote portrait alive - generating expressive portrait videos with Audio2Video diffusion model under weak conditions. In *ECCV*, 2024.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. FantasyTalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025.

Huawei Wei, Zejun Yang, and Zhisheng Wang. AniPortrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved Distribution Matching Distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Steven Xiao, Xindi Zhang, Dechao Meng, Qi Wang, Peng Zhang, and Bang Zhang. Knot Forcing: Taming autoregressive video diffusion models for real-time infinite interactive portrait animation. *arXiv preprint arXiv:2512.21734*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T Freeman. Improved Distribution Matching Distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with Distribution Matching Distillation. In *CVPR*, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. SpeakerVid-5M: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025.

Dingcheng Zhen, Shunshun Yin, Shiyang Qin, Hou Yi, Ziwei Zhang, Siyuan Liu, Gan Qi, and Ming Tao. Teller: Real-time streaming audio-driven portrait animation with autoregressive motion generation. In *CVPR*, pp. 21075–21085, 2025. doi: 10.1109/CVPR52734.2025.01963.

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. SoulX-LiveAct: Towards hour-scale real-time human animation with Neighbor Forcing and ConvKV memory. *arXiv preprint arXiv:2603.11746*, 2026.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeItTalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6): 221:1–221:15, 2020. doi: 10.1145/3414685.3417774.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal Forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *ICML*, 2026.

A WHY THE NOISE LEVEL MATTERS FOR DFD: AN IDEALIZED ANALYSIS

This section analyzes, under idealized assumptions, how the DFD update changes with the noise level at which the scores are queried. The analysis motivates noise-stage routing (§3.4). Proposition 1 describes what the real trajectory adds to the DMD update at high noise, and Proposition 2 shows that the same supervision approaches a regression onto the paired real video at low noise. Propositions 3 and 4 compare DFD and DMD when the student has collapsed onto a static mode of the teacher.

Setup. We fix the conditioning signal c and assume that the teacher distribution equals the real data distribution. The teacher and the fake score are modeled as exact conditional-mean denoisers $m_{T,t}$ and $m_{F,t}$ for the teacher distribution and the current student distribution, respectively; each maps a noisy observation to the conditional expectation of the clean latent. For analytical tractability, the teacher is modeled without classifier-free guidance. All distributions have finite second moments, and we write $\mu_c := \mathbb{E}_T[X \mid c]$ and $\Sigma_c := \mathrm{Cov}_T(X \mid c)$ for the conditional mean and covariance of the teacher. The student output is $x^s = G_\theta(\xi, c)$ for a random seed ξ, and, given c, the seed ξ, the real video x^r, and the shared standard Gaussian noise ϵ are mutually independent. Expectations over x^r treat it as a draw from the conditional data distribution; since the training set contains one real video per condition, these expectations describe a population-level model. We analyze the update after dividing out the time weight $w_t > 0$ and before region masking.

Notation. Following the shared-noise coupling in §3.4 (Eq. 5), we write $\delta = x^s - x^r$ and $r_t := \mathrm{SNR}_t = \alpha_t^2/\sigma_t^2$ with $\alpha_t, \sigma_t > 0$, and define the unweighted updates $u_{\mathrm{DFD}} = g_{\mathrm{DFD}}/w_t$ and $u_{\mathrm{DMD}} = g_{\mathrm{DMD}}/w_t$. Here t is the level at which the clean student rollout is re-noised for the score queries, which is distinct from the sampling steps of the student. Since these updates are gradients, the student moves against them; a term of the form $K(x^s - y)$ with $K \succeq 0$ therefore pulls x^s toward y.

A.1 COVARIANCE-WEIGHTED CORRECTION AT HIGH NOISE

DFD differs from DMD only in the input of the teacher query, so the two updates share the fake-score term.

Proposition 1 (Covariance-weighted correction)**.** *Under the setup above, the additional term that DFD introduces over DMD,*

$$b_t := u_{\mathrm{DFD}} - u_{\mathrm{DMD}} = m_{T,t}(z_t^s) - m_{T,t}(z_t^r), \tag{8}$$

satisfies

$$b_t = A_t \delta, \qquad A_t := r_t \int_0^1 C_t\big(z_t^r + s\,\alpha_t \delta\big)\, ds \succeq 0, \tag{9}$$

where $C_t(z) := \mathrm{Cov}_T(X \mid Z_t = z, c)$ is the posterior covariance of the teacher. For a fixed pair (x^s, x^r) and noise ϵ, $A_t = r_t \Sigma_c + o(r_t)$ as $r_t \to 0$.

Proof. Under the Gaussian observation model $Z_t = \alpha_t X + \sigma_t \epsilon$, the Jacobian of the conditional mean satisfies the conditional covariance identity $\nabla_z m_{T,t}(z) = (\alpha_t/\sigma_t^2)\, C_t(z)$ (Dytso et al., 2023). Since $z_t^s - z_t^r = \alpha_t \delta$, integrating this identity along the segment between the two points gives Eq. 9, and $A_t \succeq 0$ because every conditional covariance is positive semidefinite. As $r_t \to 0$ with x^s, x^r, and ϵ fixed, the posterior of X given Z_t converges to the conditional prior, and dominated convergence with finite second moments gives $\int_0^1 C_t\big(z_t^r + s\,\alpha_t \delta\big)\, ds \to \Sigma_c$. □

If the posterior covariance is bounded along the segment, $\|C_t(z)\|_{\mathrm{op}} \le K_c$, then $\|b_t\| \le K_c\, r_t \|\delta\|$, so the correction vanishes as $r_t \to 0$. Because A_t is a matrix, the correction also depends on direction. In the high-noise limit, its leading-order action is governed by the eigenspectrum of Σ_c: directions of larger conditional variance receive larger gains, and zero-variance directions receive none. Since Σ_c is the covariance of the teacher, this weighting reflects the variation that the data admit under c, including variation that the student does not currently produce. At a finite noise level, A_t averages the posterior covariance along the segment and need not share the eigenvectors of Σ_c.

Gaussian example. The following example gives a closed form in which the role of the spectrum can be read off; we do not assume elsewhere that video latents are Gaussian. If $X \mid c \sim \mathcal{N}(\mu_c, \Sigma_c)$

with $\Sigma_c = U\,\mathrm{diag}(\lambda_i)\,U^\top$, the posterior covariance does not depend on z, $C_t = \Sigma_c(I + r_t\Sigma_c)^{-1}$, and $A_t = r_t\Sigma_c(I + r_t\Sigma_c)^{-1}$. Projecting b_t onto the i-th eigenvector u_i gives $u_i^\top b_t = h_i(t)\,u_i^\top\delta$ with

$$h_i(t) = \frac{r_t\lambda_i}{1 + r_t\lambda_i} \in [0, 1). \tag{10}$$

We call $h_i(t)$ the *transfer gain* of direction u_i: the fraction of the trajectory difference along u_i that DFD adds to the DMD update. The gain increases with λ_i at every noise level, and for $\lambda_i > \lambda_j > 0$ the ratio $h_i(t)/h_j(t) = \lambda_i(1+r_t\lambda_j)/[\lambda_j(1+r_t\lambda_i)]$ decreases monotonically from λ_i/λ_j as $r_t \to 0$ to 1 as $r_t \to \infty$. A lower SNR therefore makes the transfer more selective and weaker in every direction, and a higher SNR raises the gain of every positive-variance direction toward one.

A conditional noise window. Suppose that, once the reference frame, audio, and text are fixed, the motion to be recovered lies mainly in directions with a large conditional variance $\lambda_{\mathrm{motion}}$, whereas frame-specific details that should not be copied from the paired video lie in directions with a small conditional variance $\lambda_{\mathrm{detail}}$. If $\lambda_{\mathrm{motion}} \gg \lambda_{\mathrm{detail}}$, then in the range

$$\frac{1}{\lambda_{\mathrm{motion}}} \lesssim r_t \ll \frac{1}{\lambda_{\mathrm{detail}}} \tag{11}$$

the motion gain is of order one, with $h_{\mathrm{motion}} \geq 1/2$ once $r_t\lambda_{\mathrm{motion}} \geq 1$, while $h_{\mathrm{detail}} \approx r_t\lambda_{\mathrm{detail}} \ll 1$. Within this window, DFD transfers the motion difference between the real and generated videos with an order-one gain and strongly suppresses the cross-trajectory pull along detail directions. The correspondence between high-variance directions and motion is a task-level assumption. The observation in §3.2 that the non-mouth person region admits many plausible head motions, expressions, and gestures for the same speech input is consistent with this assumption.

A.2 Cross-Trajectory Regression at Low Noise

At low noise, the full DFD update approaches a regression onto the paired real video, for general teacher and student distributions.

Proposition 2 (Low-noise regression)**.** *Under the setup above, let d be the latent dimensionality. Then*

$$\mathbb{E}\big\|u_{\mathrm{DFD}} - (x^s - x^r)\big\|^2 \leq \frac{4d}{r_t}, \tag{12}$$

so $u_{\mathrm{DFD}} \to x^s - x^r$ in mean square as $r_t \to \infty$. If, in addition, the generator Jacobian $J_\theta := \partial G_\theta(\xi, c)/\partial\theta$ satisfies $\|J_\theta\|_{\mathrm{op}} \leq K_J$, then the expected parameter update $\mathbb{E}[J_\theta^\top u_{\mathrm{DFD}}]$ differs by at most $2K_J\sqrt{d/r_t}$ from the gradient of the paired regression objective

$$\mathcal{L}_{\mathrm{pair}}(\theta \mid c) = \tfrac{1}{2}\,\mathbb{E}_{\xi,x^r|c}\|G_\theta(\xi, c) - x^r\|^2 = \tfrac{1}{2}\,\mathbb{E}_{\xi|c}\|G_\theta(\xi, c) - \mu_c\|^2 + \tfrac{1}{2}\,\mathrm{tr}\,\Sigma_c. \tag{13}$$

Proof. The conditional mean minimizes the mean-squared reconstruction error, so its error is at most d/r_t, the error of the naive estimator Z_t/α_t. Applying this bound to the student term $m_{F,t}(z_t^s) - x^s$ and to the teacher term $m_{T,t}(z_t^r) - x^r$, and using $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, gives Eq. 12. The gradient of $\mathcal{L}_{\mathrm{pair}}$ is $\mathbb{E}[J_\theta^\top(x^s - x^r)]$, so Jensen's inequality and Eq. 12 bound the difference by $K_J\sqrt{4d/r_t}$. The second form of $\mathcal{L}_{\mathrm{pair}}$ follows from $\mu_c = \mathbb{E}[x^r \mid c]$ and the conditional independence of ξ and x^r. □

The objective $\mathcal{L}_{\mathrm{pair}}$ pulls outputs from all seeds toward the same conditional mean. When matched real videos differ in hand positions, head angles, or body poses, this cross-trajectory averaging can blur structural details in individual generations. The bound in Eq. 12 grows with d and characterizes the limit $r_t \to \infty$.

A.3 A Teacher with Static and Dynamic Modes

To compare the teacher queries of DFD and DMD when the student has collapsed onto a static video, we consider a teacher with one static mode and two dynamic modes. Let

$$F(\rho, \varphi) = \big(\rho,\ \ell\cos\varphi,\ \ell\sin\varphi\big) \in \mathbb{R}^3,$$

where ρ measures how far a video moves away from the static video and φ selects the action. The teacher places mass w_0 on the static video $x_0 = F(0,0)$ and mass $w_\pm = (1-w_0)/2$ on each dynamic video $x_\pm = F(D, \pm\varphi_*)$ with $0 < \varphi_* < \pi/2$, and the real video is fixed to $x^r = x_+$. The student is

$$G_\theta(\xi) = F(\bar\rho + v\xi,\ \bar\varphi + a\xi), \qquad \Pr(\xi = 1) = \Pr(\xi = -1) = \tfrac{1}{2},$$

and all four parameters $\theta = (\bar\rho, v, \bar\varphi, a)$ are trained: $\bar\rho$ and $\bar\varphi$ move both seeds together, and v and a separate them. In particular, the student can place both seeds on the same dynamic video, $\theta = (D, 0, \varphi_*, 0)$. We measure the motion relative to the static video and the variation across seeds by

$$\begin{aligned} \mathcal{A}(\theta) &= \mathbb{E}_\xi \|G_\theta(\xi) - x_0\|^2 = \bar\rho^2 + v^2 + 2\ell^2(1 - \cos a \cos\bar\varphi), \\ \mathcal{V}(\theta) &= \operatorname{tr}\operatorname{Cov}_\xi G_\theta(\xi) = v^2 + \ell^2 \sin^2 a. \end{aligned} \tag{14}$$

Each dynamic video has $\mathcal{A} = D^2 + 2\ell^2(1 - \cos\varphi_*)$, and a student that places its two seeds on x_+ and x_- has $\mathcal{V} = \ell^2 \sin^2\varphi_*$.

Which modes each query returns. Write $y = z_t/\alpha_t = x + \epsilon/\sqrt{r}$ with $r = r_t$. The teacher posterior assigns weight $p_j(y) \propto w_j \exp(-r\|y - x_j\|^2/2)$ to x_j, and $m_{T,t} = \sum_j p_j x_j$. Let $\Delta^2 = \|x_\pm - x_0\|^2 = D^2 + 2\ell^2(1 - \cos\varphi_*)$ and $L_{jk}(y) = w_j e^{-r\|y-x_j\|^2/2} / \big(w_k e^{-r\|y-x_k\|^2/2}\big)$. Since $p_j \le \min\{1, L_{jk}\} \le \sqrt{L_{jk}}$, and $\mathbb{E}_\epsilon \sqrt{L_{jk}(x_k + \epsilon/\sqrt{r})} = \sqrt{w_j/w_k}\, e^{-r\|x_j - x_k\|^2/8}$,

$$\mathbb{E}_\epsilon\big[p_+ + p_-\big](x_0 + \epsilon/\sqrt{r}) \le \sqrt{\tfrac{2(1-w_0)}{w_0}}\, e^{-r\Delta^2/8}, \qquad \mathbb{E}_\epsilon\, p_0(x_+ + \epsilon/\sqrt{r}) \le \sqrt{\tfrac{2w_0}{1-w_0}}\, e^{-r\Delta^2/8}. \tag{15}$$

When $rD^2 \gg 1$, the query at a static student therefore returns the static mode, and the query at the real video returns the dynamic modes. How the second query divides its weight between x_+ and x_- depends on $r\ell^2$, and for $r\ell^2 \ll 1$ each receives close to one half. In the notation of Eq. 11, D plays the role of the motion scale and ℓ that of the detail scale. With $(D, \ell, \varphi_*, w_0, r) = (25, 1, \pi/3, 0.1, 0.1)$, so that $rD^2 = 62.5$ and $r\ell^2 = 0.1$, the query at x_0 gives the two dynamic modes a total weight of 1.7×10^{-4}, and the query at x_+ gives x_+ and x_- the weights 0.535 and 0.465.

Proposition 3 (DFD fixed point)**.** *Let $\tau = \mathbb{E}_\epsilon\, m_{T,t}(x^r + \epsilon/\sqrt{r}) = (\tau_\rho,\ M\cos\beta,\ M\sin\beta)$ and $B(\kappa) = \mathbb{E}_{Z\sim\mathcal{N}(0,1)} \tanh(\kappa + \sqrt{\kappa}\, Z)$. With an exact fake score, the unweighted DFD population gradient is*

$$\begin{aligned} g_{\bar\rho} &= \bar\rho - \tau_\rho, & g_v &= v\, B(\kappa), \\ g_{\bar\varphi} &= \ell M \cos a\, \sin(\bar\varphi - \beta), & g_a &= \ell \sin a\, \big[M\cos(\bar\varphi - \beta) - \ell\,(1 - B(\kappa))\cos a\big], \end{aligned} \tag{16}$$

where $\kappa = r(v^2 + \ell^2 \sin^2 a)$. If $0 < M < \ell$, the gradient flow $\dot\theta = -g$ has the locally stable fixed point $\theta^* = (\tau_\rho, 0, \beta, a^*)$, where $a^* \in (0, \pi/2)$ is the unique solution of $[1 - B(r\ell^2 \sin^2 a^*)]\cos a^* = M/\ell$.

Proof. The student has two equally likely outputs with midpoint $\bar x$ and half difference $\hat x$, where $\|\hat x\|^2 = v^2 + \ell^2 \sin^2 a$, so its posterior mean, as a function of y, is $\bar x + \hat x \tanh(r\langle y - \bar x, \hat x\rangle)$. At $y = G_\theta(\xi) + \epsilon/\sqrt{r}$, averaging over ϵ gives $\bar x + \xi \hat x\, B(r\|\hat x\|^2)$. The teacher term averages to τ because ξ and ϵ are independent. Multiplying the difference by $\partial G_\theta(\xi)/\partial\theta$ and averaging over ξ gives Eq. 16. At a fixed point with $0 < a < \pi/2$, $g_{\bar\rho} = g_v = 0$ give $\bar\rho = \tau_\rho$ and $v = 0$, and $g_{\bar\varphi} = g_a = 0$ give $\bar\varphi = \beta$ together with the equation for a^*, since $\bar\varphi = \beta + \pi$ would make the bracket in g_a negative. The left-hand side of that equation decreases strictly from 1 to 0 on $(0, \pi/2)$. The Jacobian of g at θ^* is diagonal with the positive entries 1, $B(r\ell^2 \sin^2 a^*)$, $\ell M \cos a^*$, and $-\ell^2 \sin a^* \,\frac{\mathrm{d}}{\mathrm{d}a}\big\{[1 - B(r\ell^2 \sin^2 a)]\cos a\big\}\big|_{a=a^*}$. □

For the exact teacher, the angular part of τ is a convex combination of distinct points on the circle of radius ℓ, so $M < \ell$ at every finite r, and the stable fixed point places the two seeds on the distinct dynamic videos $F(\tau_\rho, \beta \pm a^*)$. Seed differences grow even after the common action matches the target: at $v = 0$ and $\bar\varphi = \beta$, $\dot a = \ell(\ell - M)\, a + o(a)$. Up to the static weight bounded in Eq. 15, $\tau = \big(D,\ \ell\cos\varphi_*,\ \ell\sin\varphi_*\, B(r\ell^2 \sin^2\varphi_*)\big)$. When $rD^2 \to \infty$ and $r\ell^2 \to 0$, $\tau \to (D, \ell\cos\varphi_*, 0)$, so $\theta^* \to (D, 0, 0, \varphi_*)$ and the two seeds approach x_+ and x_-. As $r \to \infty$ instead, the query at x^r resolves the dynamic mode, $\tau \to x_+$, $M \to \ell$, and $a^* \to 0$, so both seeds copy the real video, in line with Proposition 2.

Proposition 4 (DMD invariant region)**.** *Let* $(D, \ell, \varphi_*, w_0, r) = (25, 1, \pi/3, 0.1, 0.1)$. *The box* $\mathcal{U} = \{|\bar{\rho} \pm v| \le 0.6,\ |\bar{\varphi} \pm a| \le 0.2\}$ *is invariant under the flow of the unweighted DMD population gradient at this noise level. Along every DMD trajectory that starts in* $\mathcal{U}$, $\mathcal{A} \le 0.36 + 2(1 - \cos 0.2) < 0.40$ *and* $\mathcal{V} \le 0.36 + \sin^2 0.2 < 0.40$.

Proof. DMD has the same fake term as DFD and replaces τ by the teacher output at the student sample, $\bar{T}(\xi) = \mathbb{E}_\epsilon\, m_{T,t}(G_\theta(\xi) + \epsilon/\sqrt{r}) = x_0 + \sum_{j=\pm} e_j(\xi)(x_j - x_0)$ with $e_j(\xi) = \mathbb{E}_\epsilon\, p_j$. For $j = \pm$, $p_j \le \min\{1, L_{j0}\}$, and $\log L_{j0}$ is Gaussian in ϵ with variance $r\Delta^2$ and a mean that is affine in $G_\theta(\xi)$, so $\mathbb{E}_\epsilon \min\{1, L_{j0}\}$ has a closed form that increases with this mean. Over the outputs allowed by $\mathcal{U}$, the mean is largest at $F(0.6, \pm 0.2)$, and evaluating the closed form there gives $e(\xi) := e_+(\xi) + e_-(\xi) \le 6.5 \times 10^{-4}$. For the seed $\xi = 1$, with $\rho_+ = \bar{\rho} + v$ and $\varphi_+ = \bar{\varphi} + a$, the flow gives

$$\dot{\rho}_+ = -\big(\bar{\rho} + vB(\kappa)\big) + D\, e(1),$$
$$\dot{\varphi}_+ = -\ell^2\big[\sin(\bar{\varphi} + a) - \tfrac{1}{2}\sin 2a\,(1 - B(\kappa))\big] + \nu, \qquad |\nu| \le 2\ell^2 \sin(\varphi_*/2)\, e(1).$$

For a symmetric binary input at SNR κ, the minimum mean-squared error equals $1 - B(\kappa)$ and is at most the error $1/(1+\kappa)$ of the linear estimator, so $B(\kappa) \ge \kappa/(1+\kappa)$. On the face $\rho_+ = 0.6$, $v \in [0, 0.6]$ and $\kappa \ge rv^2$, so $\bar{\rho} + vB(\kappa) = 0.6 - v + vB(\kappa) \ge 0.6 - v/(1 + rv^2) \ge 0.6 \cdot 0.036/1.036 > 0.020$, while $D\, e(1) < 0.017$. On the face $\varphi_+ = 0.2$, the bracket is at least $\sin 0.2\,(1 - \cos 0.2) > 0.0039$, while $2\ell^2 \sin(\varphi_*/2)\, e(1) < 0.00065$. On the face $\rho_+ = -0.6$, the same bound holds with the opposite sign, and the teacher term $D\, e(1) \ge 0$ also points inward; the face $\varphi_+ = -0.2$ and the faces of the seed $\xi = -1$ follow by symmetry, so the flow points into $\mathcal{U}$ on its whole boundary. □

Comparison at the same noise level. For the parameters of Proposition 4, $\tau = (24.9995, 0.5000, 0.0606)$, so $M = 0.504 < \ell$ and Proposition 3 gives $\theta^* = (24.9995, 0, 0.121, 1.001)$. From the same initialization $\theta_0 = (0.01, 0.02, 0.01, 0.02) \in \mathcal{U}$, the DFD flow converges to θ^*, and the DMD flow settles at $(0.0045, 0.227, 0, 0)$. Near the static video, the query at the student returns the static mode (Eq. 15), so DMD receives almost no supervision toward the dynamic modes. The query at the real video returns both dynamic modes, so τ carries their motion, and its angular part lies strictly inside the circle traced by F. DFD therefore moves the common motion to τ_ρ and separates the seeds between the two actions. Table 4 also lists the DFD fixed point at $r = 10$: it has almost the same $\mathcal{A}$ as at $r = 0.1$ and a much smaller $\mathcal{V}$, because the query then resolves the dynamic mode and both seeds copy the real video.

Table 4: **Motion and seed variation in the model of Appendix A.3.** The parameters are $(D, \ell, \varphi_*, w_0) = (25, 1, \pi/3, 0.1)$, and $\mathcal{A}$ and $\mathcal{V}$ follow Eq. 14. Each dynamic video of the teacher has $\mathcal{A} = 626$, and seeds placed on x_+ and x_- give $\mathcal{V} = 0.75$.

Update	r	$\mathcal{A}$	$\mathcal{V}$
DMD, from θ_0	0.1	0.052	0.052
DFD, fixed point	0.1	625.90	0.709
DFD, fixed point	10	625.99	0.0007

A.4 Implications for Noise Routing

The analysis separates three regimes of the query noise level. (i) As $r_t \to 0$, the correction b_t of Proposition 1 vanishes, and the unweighted DFD update approaches the DMD update. (ii) At finite high noise, DFD adds a covariance-weighted correction to DMD. Under the spectral conditions of Eq. 11, this correction transfers motion differences with an order-one gain and suppresses detail differences. In the model of §A.3, where the student starts near a static mode of the teacher, the query at the real video returns the dynamic modes, so DFD moves the student to them and separates its seeds between two actions, while DMD at the same noise level stays near the static mode. (iii) At low noise, DFD approaches the paired regression of Proposition 2, which pulls all seeds toward the

same mean and toward the frame-level layout of a different real video; in the model of §A.3, both seeds then copy the real video.

NR-DFD (Eq. 6) is designed to use regime (ii) and to leave regime (iii) to DMD, which refines details along the student's own trajectory. Its single threshold also includes the highest noise levels of regime (i), where the unweighted difference between DFD and DMD vanishes. The common intuition that high noise levels govern coarse structure and low noise levels govern details applies to these query levels as a statement about which variations each objective emphasizes. Whether conditional diversity improves in practice is an empirical question, which we address with the multi-seed diversity and temporal dynamics metrics in Tables 1 and 3.

B Mask Analysis

Table 5: **Mask alignment between real video and student rollout video.** Person masks are extracted via a segmentation model, and mouth masks via facial landmarks.

Person coverage (%) ↑	Person IoU (%) ↑	Mouth coverage (%) ↑	Mouth IoU (%) ↑
94.49	89.73	43.61	32.74

Table 6: **Mouth mask source ablation.** Comparison of training with mouth masks from the student rollout video versus ground-truth video.

Mouth mask source	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	Step time (s) ↓
Student rollout	**25.43**	309.92	**4.07**	**10.94**	63.73
Real video	25.49	**307.91**	4.03	10.99	**61.11**



Figure 5: **Comparison of mouth masks between real video and student rollout video.**

To determine whether masks should be extracted from the student rollout video or the real video, we conduct an ablation study. Table 5 reports the mask alignment between real videos and student rollout videos: person-region coverage reaches 94.49% with an IoU of 89.73%, indicating highly consistent person placement, so we directly use person masks from real videos. Mouth coverage is only 43.61% with an IoU of 32.74%. As shown in Figure 5, mouth masks extracted from real videos exhibit noticeable misalignment with the actual mouth position in student rollout videos when the subject turns their head. We further compare training with mouth masks from the student rollout video versus the real video. As shown in Table 6, using student-rollout mouth masks yields better Sync-C and Sync-D, with negligible additional time overhead (63.73 vs. 61.11 s/step). We therefore extract mouth masks from the student rollout video.

C More Implementation Details

Routed Forcing training uses the following setup: both the student and fake score use the Adam optimizer with $\beta_1 = 0.0$, $\beta_2 = 0.999$, no weight decay, no learning rate warmup or decay. The student learning rate is 1×10^{-6} and the fake score learning rate is 2×10^{-6}. The per-GPU batch size is 1 on 8 GPUs with gradient accumulation of 2, yielding an effective batch size of 16. The student updates once every 5 fake score updates (1:5 ratio), i.e., the fake score trains at 5× the frequency of

the student. The classifier-free guidance scale is 4.5. We sample $u \sim \mathcal{U}(0, 1)$ and obtain the score timestep using the shift-5 schedule $t = 5u/(1 + 4u)$. "Top-50%" denotes the upper half of this shifted timestep distribution, implemented as $t \geq 0.832$.

For dynamics evaluation, we exclude the conditioning frame and sample at 8 FPS. RAFT-Motion averages the largest 5% of optical-flow magnitudes between consecutive sampled frames, normalized by the image diagonal and temporal interval. DINO-Temporal is the mean cosine distance between DINOv2 features of consecutive sampled frames.

D Additional Experimental Results

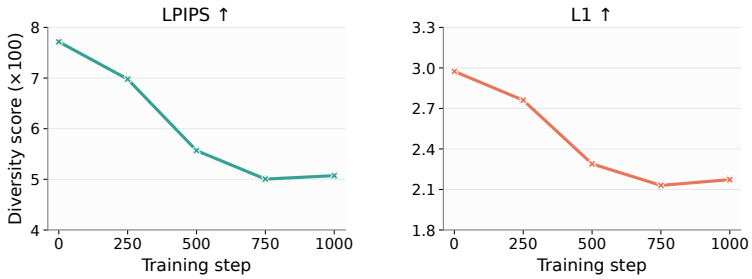


Figure 6: **Diversity decreases during DMD training.** Starting from the LiveTalk checkpoint, we train Self Forcing with DMD for 1,000 steps and measure diversity (mean pairwise LPIPS and L_1 across four seeds over 100 SpeakerVid samples) at each checkpoint. Both metrics drop during training, confirming that DMD training itself drives diversity collapse.

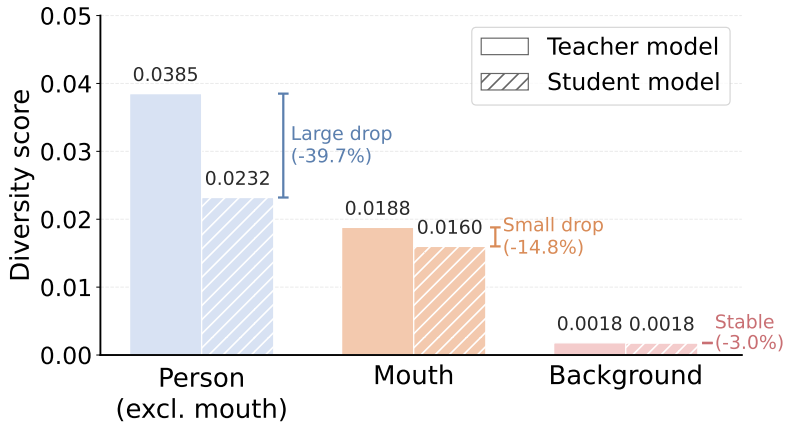


Figure 7: **Regional diversity measured by L_1.** Mean pairwise L_1 distance across different random seeds for teacher and student models, confirming the same trend as the LPIPS-based analysis in §3.2: the person region suffers the largest diversity drop (-39.7%), the mouth region shows a moderate drop (-14.8%), and the background remains stable (-3.0%).

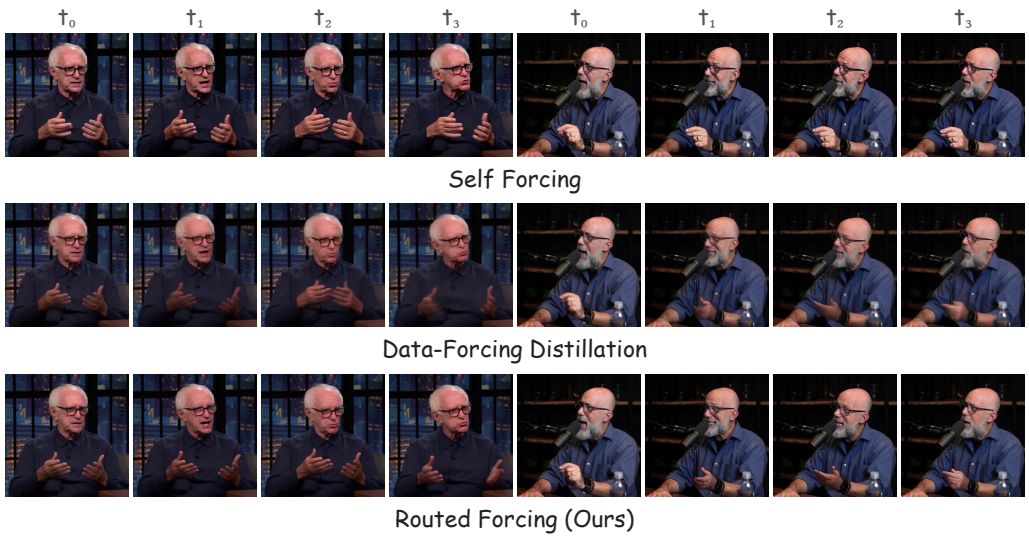


Figure 8: **Additional qualitative comparison of training strategies.** Each column shows frames at the same timestep across methods. Routed Forcing produces more dynamic and diverse motion than Self Forcing while maintaining better visual quality than DFD.