LOST IN THE ROLLOUT: RETHINKING ON-POLICY DISTILLATION FOR MULTIMODAL VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation, exemplified by Self Forcing, has become the leading approach for turning video diffusion models into fast, interactive few-step generators. Applying it under joint text, image, and audio conditioning for avatar generation has recently attracted great interest, but rare insights have been exposed on the corresponding training recipe. This paper first identifies the *lost in the rollout* issue in this regime, where the video rollouts generated by the student often lose what they should carry at *two levels*. *First*, they lose the conditions: within a few hundred steps, training collapses into black frames, flickering, or blur, and the identity given by the image condition is gone. Unlike a text prompt, the image condition becomes the first frame of every rollout, and the teacher and the critic score each rollout under that same image, so a flaw in the image enters the very signal meant to correct the student. (1) Curated conditions and (2) a converged ODE initialization keep the conditions intact and stabilize training; together they cut FID from 27.10 to 11.67 and raise Sync-C from 3.13 to 4.15. *Second*, they lose the teacher's trajectory: after the ODE initialization, the student sees only the DMD score of its own rollouts, and at standard learning rates and guidance it learns too little lip-sync. We keep a small piece of the regression onto teacher samples that DMD2 removed: (3) an auxiliary target that matches the audio energy of each generated clip to that of a full teacher sample for the same condition. Because the target is computed on the generated audio, it applies to any model that generates sound. (4) Higher learning rates and teacher guidance make the student learn more from the teacher's score, raising Sync-C further to 4.50. With these remedies, a collapsing recipe becomes a stable four-step causal student of a Wan2.1-based multimodal model, which improves lip-sync, image quality, and aesthetics over the bidirectional model of the same size on HDTF (Zhang et al., 2021), AVSpeech (Ephrat et al., 2018), and CelebV-HQ (Zhu et al., 2022), with over 20× higher throughput, running at 24.82 FPS on a single GPU. On MiniMax-H3, a 33B model that generates video and audio jointly, the auxiliary target further sharpens lip-sync, raising Sync-C from 7.37 to 8.02. The lesson reaches beyond these models: an on-policy student excels when its rollouts start well and the teacher's signal keeps reaching them.

1 INTRODUCTION

Diffusion transformers (DiTs) (Peebles & Xie, 2023; Brooks et al., 2024; Team Wan et al., 2025; Chen et al., 2025) have enabled appealing visual fidelity for video generation. The sampling involves denoising all video frames simultaneously using bidirectional attention via an iterative process. It can be costly, e.g., models such as Seedance (Gao et al., 2025), MiniMax-H3 (MiniMax, 2026), Veo (Google DeepMind, 2025) require about 1 to 2 minutes to generate a single 5 to 10 second clip, incurring prohibitive inference expense and creating fundamental barriers to real-time applications.

As a result, there is increasing interest in converting pretrained, bidirectional, many-step video diffusion models into causal, few-step autoregressive (AR) ones via distillation techniques (Lu & Thinking Machines Lab, 2025; Agarwal et al., 2024; Yin et al., 2024b; Huang et al., 2025a; Liu et al., 2025b). The resultant models enable low-latency streaming generation. Recent real-time audio-driven systems distill such models as part of their recipes (Low & Wang, 2025; Huang et al., 2025b;

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.

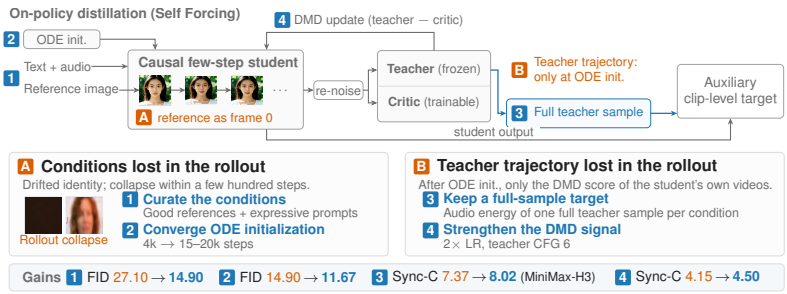

Figure 1: **What on-policy distillation under multimodal conditioning loses in the rollout, and how we keep it.** Top: the Self Forcing loop and where the conditions and the teacher's trajectory are lost. Bottom left: the image condition enters the rollout as its first frame, and with it the training loop: every generated frame depends on it, so its flaws are carried into the student's training (frames from Fig. 2, step 500); curated conditions (1) and a converged ODE initialization (2) stabilize it (Fig. 3c,d). Bottom right: after the ODE initialization the teacher's trajectory no longer reaches the student; a clip-level target from one full teacher sample per condition brings a piece of it back and improves lip-sync on MiniMax-H3 (3; Tab. 2); doubled learning rates with teacher CFG 6 strengthen the remaining signal (4; Fig. 3g). Bottom row: the gain of each remedy (Tab. 3; (3): Tab. 2).

Shen et al., 2025; Zhang et al., 2026). These works, however, treat distillation as one component of a real-time system; why the distillation itself breaks under the simultaneous integration of text, images, and audio, as in audio-driven human video generation, remains underexplored. Such natural conditioning is critical for building general-purpose interactive AI systems that can not only *understand* across modalities but also *express themselves visually* for natural human-AI interaction.

This paper studies on-policy distillation under such multimodal conditioning, starting from its leading approach, Self Forcing (Huang et al., 2025a). Self Forcing initializes a few-step causal student by ODE trajectory distillation (Song et al., 2023; Yin et al., 2025) and then trains it on its own rollouts with distribution matching distillation (DMD) (Yin et al., 2024b;a), so that the student learns from the videos it actually generates and *exposure bias* (Ning et al., 2023; Schmidt, 2019) is reduced. Applying it to a multimodal video diffusion model therefore *seems* straightforward. However, naively applying Self Forcing training produces extensive visual artifacts, e.g., flickering (row 1 of Fig. 5), and a related issue has been reported in the community (BrianChen1129, 2025). The failure is unexpected: on text-to-video, Self Forcing succeeds even from ODE checkpoints that generate low-quality videos (Xilluill, 2025). We term this phenomenon *lost in the rollout*. Trained on its own rollouts, the student loses two things in them. First, it loses the conditions. A text prompt instructs the model what to generate, and the teacher's score corrects a poor student video; an image condition instead supplies what the video is made of, as the first frame of every rollout shared by the teacher and the critic, so its flaws enter the very signal meant to correct the student (Section 3.2). Second, it loses the teacher's trajectory: after the ODE initialization the student no longer sees what the teacher generates, only the DMD score of its own rollouts.

We counter each loss with targeted improvements, which together form our distillation recipe. To keep the *conditions* in the rollout: (1) **refining multimodal conditions for distillation**, e.g., curating the image and text conditions that the training loop feeds back to the student; and (2) **training ODE initialization to convergence** before applying on-policy DMD training. To keep the *teacher's signal*: (3) **keeping an auxiliary audio-energy target from a full teacher sample**, a small piece of the regression onto teacher samples that DMD (Yin et al., 2024b) used and DMD2 (Yin et al., 2024a) removed; and (4) **strengthening what the student learns from the teacher's score** through aggressive learning rates and tuned classifier-free guidance (CFG) (Ho & Salimans, 2022) scales. We validate these by distilling a multimodal variant of Wan2.1 (Team Wan et al., 2025) (i.e., OmniAvatar (Gan et al., 2025)) and benchmarking on HDTF (Zhang et al., 2021), AVSpeech (Ephrat et al., 2018), and CelebV-HQ (Zhu et al., 2022); we test (3) on MiniMax-H3 (MiniMax, 2026), a 33B model that generates video and audio jointly (Section 4.3). We show that an on-policy student improves when its rollouts start from faithful conditions and the teacher's signal keeps reaching them.

In summary, our contributions are:

- **Identifying the lost-in-the-rollout phenomenon.** We show that on-policy distillation under multimodal conditioning fails through the student's own rollouts: the image condition becomes the first frame of every rollout, so its flaws enter the student's training signal, and after the ODE initialization the rollout also replaces the teacher's trajectory.

- **Recovering what the rollout loses.** (1) Curated multimodal conditions and (2) a converged ODE initialization keep the conditions and stabilize training. (3) A target that matches the audio energy of each generated clip to that of a full teacher sample brings back a piece of the teacher's trajectory and raises Sync-C by 0.65 on MiniMax-H3, and (4) higher learning rates and teacher CFG raise Sync-C from 4.15 to 4.50 on the Wan2.1-based model.

- **Real-time multimodal video generation with 20× speedup.** Our distilled 4-step model achieves real-time generation at 24.82 FPS on a single GPU, improving lip-sync, image quality, and aesthetics over the bidirectional model of the same size (Tab. 1).

2 Related Work

Multimodal Video Diffusion. Modern video synthesis has evolved beyond text control to incorporate richer conditioning signals such as images and audio (Team Wan et al., 2025; Kong et al., 2024; Chen et al., 2025; Gan et al., 2025), enabling image-guided video editing (Team Wan et al., 2025; Kong et al., 2024) and multimodal-driven virtual avatars (Gan et al., 2025; Chen et al., 2025), with work on architectural designs (Ju et al., 2025; Lin et al., 2025; Hu et al., 2025), training overheads (Team Wan et al., 2025; Kong et al., 2024), and cross-modal alignment (Li et al., 2024; Gan et al., 2025; Chen et al., 2025). These models denoise entire sequences over many steps, and their inference cost and latency render them impractical for real-time applications. Recent work (Low & Wang, 2025; Huang et al., 2025b; Shen et al., 2025; Ki et al., 2026; Liu et al., 2026; Su et al., 2026; Li et al., 2026a; Zhang et al., 2026) has begun exploring real-time audio-driven models, largely as system recipes. Recent analyses of these models attribute individual failures to individual causes, such as the audio tokens (Su et al., 2026) or the cross-modal alignment (Zhong et al., 2024). They treat the conditions as fixed inputs. We ablate the conditions and the initialization of on-policy training step by step and find that the conditions do not stay fixed: the image condition becomes the first frame of every rollout, so its flaws enter the student's training signal, and even curated conditions leave blur and flicker until the ODE initialization has converged (Section 3.2).

Real-Time Video Diffusion. Real-time video diffusion combines AR and diffusion components: AR enables streaming generation without future frame dependencies, while diffusion ensures high visual fidelity (Teng et al., 2025; Bruce et al., 2024; Liu et al., 2025b; Yang et al., 2025). AR video diffusion faces *exposure bias* (Ning et al., 2023; Schmidt, 2019; Huang et al., 2025a) from error accumulation, which post-training strategies address by distilling guidance from bidirectional models (Yin et al., 2025), increasing robustness to imperfect frames (Chen et al., 2024), and training on self-generated rollouts (Huang et al., 2025a), primarily for text-to-video. Recent analyses examine the DMD objective and its training dynamics: the structure of ODE initialization for AR distillation (Zhu et al., 2026), the dominant role of the CFG-augmented teacher term (Liu et al., 2025a), teacher-CFG bias (Li et al., 2026b), critic lag in video DMD (Wu et al., 2026), audio-CFG calibration for streaming avatars (Meng et al., 2026), and the mode-seeking bias of the reverse-KL objective in DMD (Chen et al., 2026). These analyses focus on individual terms of the objective, largely under text conditioning; a per-factor account of instability under joint text–image–audio conditioning is still missing. We study on-policy distillation under text, image, and audio conditioning: what the student's rollouts lose in this setting, and how to keep it.

3 What Is Lost in the Rollout under Multimodal Conditioning

This section briefly reviews Self Forcing (Huang et al., 2025a), a dominant distillation method for constructing real-time video diffusion, then shows what its rollout loses under multimodal conditioning and how we keep it.

3.1 PRELIMINARY: SELF FORCING

Self Forcing (Huang et al., 2025a) converts a bidirectional, many-step video diffusion model (Team Wan et al., 2025; Chen et al., 2025) into a causal, few-step AR one in two stages: an ODE initialization (Song et al., 2023; Yin et al., 2025; Gu et al., 2023; Berthelot et al., 2023) that yields a few-step causal student with block-wise causal attention, followed by on-policy distillation (Lu & Thinking Machines Lab, 2025; Agarwal et al., 2024) on self-generated rollouts with distribution matching distillation (DMD) (Yin et al., 2024b;a; 2025), which reduces the student's *exposure bias* (Ning et al., 2023; Schmidt, 2019). Let $\mathbf{x}_0$ denote the latent of the video frames under a compression variational auto-encoder (VAE) (Kingma & Welling, 2013; Rombach et al., 2022), $\mathbf{c}$ the conditions, and g_ϕ the student, which generates block by block (3 latent frames per block in our case).

ODE Initialization first performs trajectory distillation. It subsamples $k = 4$ timesteps from the teacher's $N = 48$ step denoising trajectory $\{\mathbf{x}_{t_j}\}_{j=0}^{N}$, based on which the causal student is asked to predict clean $\mathbf{x}_0$:

$$\mathcal{L}_{\text{ODE}} = \mathbb{E}_{t\sim\{t_{iN/k}\}_{i=0}^{k-1}} \left[\sum_b \left\| g_\phi(\mathbf{x}_t^b, t, \mathbf{c}) - \mathbf{x}_0^b \right\|_2^2 \right], \tag{1}$$

with the superscript b denoting the b-th block.

Distribution Matching Distillation (DMD) then addresses the exposure bias (Yin et al., 2024b;a; 2025) of the model after ODE initialization due to the training under teacher-forced trajectories. The algorithmic implementation introduces two extra models: a frozen teacher score network s_θ (Song et al., 2020; Yin et al., 2024b;a) and a trainable critic s_ψ, and the training alternates between updating the generator g_ϕ and the critic s_ψ. Specifically, the gradient for updating the student model is

$$-\mathbb{E}_{\tau,\hat{\mathbf{x}}_0,\mathbf{x}_\tau} \left[\left(s_\theta(\mathbf{x}_\tau, \tau, \mathbf{c}) - s_\psi(\mathbf{x}_\tau, \tau, \mathbf{c}) \right) \frac{\partial \hat{\mathbf{x}}_0}{\partial \phi} \right], \tag{2}$$

where $\hat{\mathbf{x}}_0 = g_\phi(\mathbf{z}, 0, \mathbf{c})$ with $\mathbf{z} \sim \mathcal{N}(0, I)$ and $\mathbf{x}_\tau$ comes from adding τ-time noise to $\hat{\mathbf{x}}_0$. For multimodal conditioning, s_θ is estimated with classifier-free guidance (Ho & Salimans, 2022) using separate scales per condition. The critic s_ψ learns to track the evolving distribution of the generator by minimizing the standard diffusion denoising objective:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_\tau \left[\left\| s_\psi(\mathbf{x}_\tau, \tau, \mathbf{c}) - \hat{\mathbf{x}}_0 \right\|_2^2 \right]. \tag{3}$$

3.2 FINDINGS AND MITIGATIONS: LOSING THE CONDITIONS AND THE TEACHER'S TRAJECTORY IN THE ROLLOUT

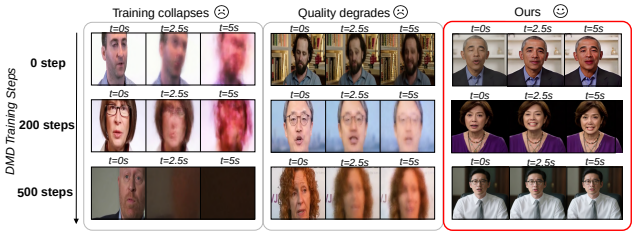

Figure 2: **Self Forcing DMD under multimodal conditioning.** Left: the default recipe on the dataset conditions degrades to black frames. Middle: conditions filtered by brightness and image quality still blur after a few hundred steps. Right: our recipe stays stable.

Applying Self Forcing (Huang et al., 2025a) with its default settings to a multimodal video diffusion model seems straightforward. However, training breaks down into visual artifacts (Fig. 2, left and middle). Why would the same recipe fail once an image and audio are added? The first clue came from the image condition: filtering the training conditions by brightness and image quality helped (Fig. 2, middle), which pointed to the image condition and to how it enters training. The answer lies in the rollout. The teacher and the critic both score the student's noised rollouts $\mathbf{x}_\tau$, and the critic score network s_ψ is trained only on these rollouts, so when the rollouts drift away from the

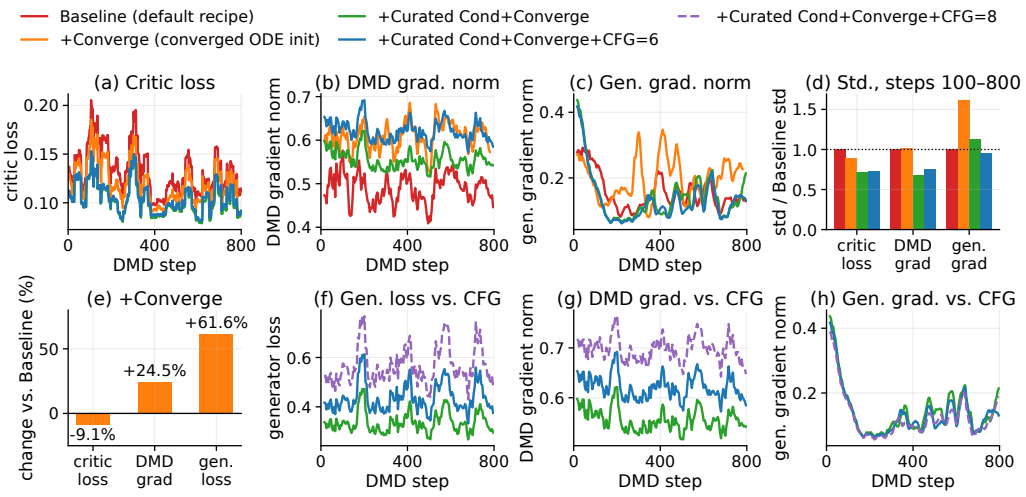

Figure 3: **Training dynamics of DMD under multimodal conditioning** (first 800 DMD steps, 30-step moving average). *Baseline*: default Self Forcing recipe with a 4k-step ODE initialization and teacher score CFG 4, on the conditions of the Tab. 3 baseline; *+Converge*: same conditions, ODE initialization trained to convergence (15k steps); *+Curated Cond+Converge*: additionally curated conditions, with a converged ODE initialization; *CFG=6/8*: *+Curated Cond+Converge* with the teacher CFG raised to 6 or 8. All configurations except *Baseline* use doubled learning rates. (a–c) The first four configurations. (d) Standard deviation of each metric over steps 100–800, relative to *Baseline*. (e) Change of the mean over steps 0–800 from *Baseline* to *+Converge*. (f–h) Generator loss, DMD gradient norm, and generator gradient norm as the teacher CFG is raised to 6 and 8.

conditions, the critic is trained on the drifted samples. Under multimodal conditioning, the conditions themselves enter the rollout: the image condition, unlike the text prompt, becomes part of it. At inference it occupies the first frame, and we write it into the first frame of the training rollout in the same way, so every later frame is generated from it, and through the teacher and the critic its flaws enter the student's own training signal. Two mitigations therefore keep the conditions in the rollout intact: curated conditions improve what the rollout starts from, and a converged ODE initialization improves what the student generates from it. The rollout also replaces the teacher's trajectory, and two further mitigations address this second loss. Fig. 3 traces how each mitigation changes the training dynamics (extended analysis in Appendix E).

Keeping the Conditions (1): Curated Multimodal Conditions. Since the image condition enters the training loop through the rollout, the conditions must be ones the loop can carry. With 2000 multimodal conditions (reference image and audio) from each of the Hallo3 and HDTF datasets and the default settings of Self Forcing, the DMD training collapses after several hundred iterations, with outputs degrading to black images (Fig. 2, left). Filtering these conditions by brightness and quality metrics still leaves blur after a few hundred steps (Fig. 2, middle), so the conditions must be improved, not only selected. Existing datasets often contain lower-quality images (such as Hallo3's overall low image quality and HDTF's facial blurriness). We therefore curate the image and text conditions $\mathbf{c}_{\text{img}}$ and $\mathbf{c}_{\text{text}}$ per dataset. For Hallo3, characterized by overall low image quality, we employ Qwen-Image (Wu et al., 2025) to generate semantically consistent yet high-quality reference frames $\mathbf{c}_{\text{img}}$. For HDTF, which primarily suffers from facial blurriness, we apply super-resolution (Zhou et al., 2022) to obtain clear facial details. We also use Qwen2.5-VL-72B (Bai et al., 2025) to refine text prompts $\mathbf{c}_{\text{text}}$, emphasizing dynamic motion and facial expressions. The configuration with curated conditions is the most stable. In Fig. 3a, *+Curated Cond+Converge* (green) is both lower and visibly smoother than *+Converge* (orange). Fig. 3d quantifies this: the standard deviation of the critic loss over steps 100–800 falls from 0.105 to 0.085. The effect is most striking in Fig. 3c, where the violent spikes of *+Converge* are almost entirely suppressed.

Keeping the Conditions (2): Converged ODE Initialization. Clean conditions, however, are not enough: even with them, blur and flicker remain (Tab. 3), because an insufficient ODE initialization

creates a weak foundation for the subsequent on-policy DMD training. We therefore train the ODE initialization to convergence, so that the student denoises well at all timesteps before DMD starts. Comparing *Baseline* (ODE 4k) with +*Converge* (ODE 15k, with doubled learning rates and teacher CFG 6), Fig. 3b,e show the combined effect of these changes: critic loss decreases by 9%, DMD gradient norm rises by 24.5%, and generator loss rises by 61.6%. Conversely, without curated conditions a converged initialization is not enough: Fig. 3c shows that +*Converge* (orange) exhibits the **most unstable** generator gradient norm, with prominent spikes at steps 300–500.

Keeping the Teacher's Signal (3): A Clip-Level Target from a Full Teacher Sample. Keeping the conditions stabilizes training, but the rollout replaces something else as well. The ODE initialization learns from the teacher's own denoising trajectories. Once DMD starts, the student sees only the teacher's and the critic's scores of its own rollouts, which tell it how its videos differ from the teacher's as a distribution. The original DMD kept a direct link to the teacher: a regression of the student's output onto teacher samples stored for the same noise (Yin et al., 2024b). DMD2 removed it together with the cost of building the paired dataset (Yin et al., 2024a), and Self Forcing follows DMD2. Is distribution matching alone enough when sound and video must stay in sync? We keep a small piece of the regression. Before training, the teacher generates one full sample for each training condition. A full regression onto this sample does not fit on-policy training: the rollout starts from its own noise, so matching the whole clip would pull the student toward one particular video. We therefore match one statistic of the whole clip instead of the clip itself. We add $\mathcal{L}_{\text{aux}} = (\log S(\hat{\mathbf{a}}_0) - \log S_{\text{ref}})^2$ to the DMD loss with weight 0.1, where $\hat{\mathbf{a}}_0$ is the generated audio latent, S is its mean energy, and S_{ref} is the same quantity on the stored teacher sample. The target is computed on the generated audio, so it applies to models that generate sound. OmniAvatar generates only video, so we evaluate the target on MiniMax-H3, which generates video and audio jointly. There it raises Sync-C over DMD alone by 0.65 (Tab. 2).

Keeping the Teacher's Signal (4): Learning More from the Teacher's Score. The teacher's score is the signal that remains, and the student must learn enough from it. With standard learning rates and guidance scales, even a stable run does not learn enough audio-visual alignment. We therefore double the learning rates and raise the teacher CFG to strengthen audio conditioning for lip-sync. Fig. 3f–h compares three CFG scales. Increasing CFG to 6 raises the DMD gradient norm by 12% (panel g), while keeping the generator loss stable (panel f) and **reducing** the generator gradient norm (panel h). One might expect a still higher scale to help further. However, the benefits do not continue: CFG=8 destabilizes training (Fig. 3f–h, purple dashed). The generator loss becomes highly unstable with substantially increased variance (panel f), and the DMD gradient norm rises by a further 12% (panel g), while the critic loss stays nearly unchanged (mean over steps 0–800: 0.105 versus 0.107): the teacher term, scaled by the CFG, dominates the DMD gradient. This gives an ordering principle: **first ensure high-quality conditions, then strong initialization, and only then amplify the learning signal**.

4 Experiments

4.1 Settings

Model and Training Configuration. We distill OmniAvatar (Gan et al., 2025), a multimodal variant of Wan2.1 (Team Wan et al., 2025) that generates videos from text, image, and audio conditions. The student is OmniAvatar-1.3B adapted to causal attention with a KV cache, generating blocks of $b = 3$ latent frames. The frozen teacher score network s_θ is OmniAvatar-14B and the trainable critic s_ψ is OmniAvatar-1.3B. We use 4000 multimodal conditions from Hallo3 (Cui et al., 2025) and the HDTF training split (Zhang et al., 2021), keeping the dataset audio and refining the image and text conditions as in Section 3.2. The student is initialized by ODE trajectory matching for 20k steps on trajectories of the bidirectional OmniAvatar-1.3B, then trained with DMD on the same conditions for 1000 steps. Hyperparameters are listed in Tab. 6. We also distill MiniMax-H3 (MiniMax, 2026), a 33B model that generates video and audio jointly from a text prompt and a first-frame image. Its student is a four-step, bidirectional rank-64 low-rank adapter on the frozen teacher (critic: a rank-128 adapter). It starts from the teacher's weights without ODE initialization and is trained with DMD for 300 updates on 240 curated HDTF conditions (512×512, 5 s). Each model uses the remedies that apply to it: OmniAvatar generates only video, so the audio-energy target (3) has nothing to act

Table 1: Quantitative comparison with existing audio-driven video generation methods (HDTF: 110 clips; AVSpeech and CelebV-HQ: 100 clips each). Our distilled model improves lip-sync (Sync-C/D), image quality (IQA), and aesthetics (ASE) over the bidirectional OmniAvatar-1.3B with over 20× higher throughput (Thr.: Throughput). OmniAvatar-14B (grey), the DMD teacher of our 1.3B student, is shown for reference and not ranked; the other methods are ranked regardless of model size. **Bold** indicates best performance and underline second best.

Methods	Model Size	Thr. (FPS) ↑	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	IQA ↑	ASE ↑
HDTF								
Ground Truth	-	-	-	-	8.06	7.31	4.20	2.56
OmniAvatar-14B (teacher) (Gan et al., 2025)	14B	0.20	8.82	151.30	6.29	9.16	3.91	2.30
AniPortrait (Wei et al., 2024)	2.5B	0.38	21.41	323.08	1.17	13.32	4.03	2.40
FantasyTalking (Wang et al., 2025)	14B	0.35	24.54	488.53	2.92	12.21	3.78	2.14
OmniAvatar-1.3B (Gan et al., 2025)	1.3B	0.97	**10.85**	**187.46**	3.85	11.38	3.87	2.25
Ours (OmniAvatar)	1.3B	**24.82**	13.68	190.07	**4.50**	**11.00**	**4.13**	**2.52**
AVSpeech								
Ground Truth	-	-	-	-	5.83	8.19	4.11	2.50
OmniAvatar-14B (teacher) (Gan et al., 2025)	14B	0.20	31.21	450.09	5.18	9.17	3.77	2.25
AniPortrait (Wei et al., 2024)	2.5B	0.38	32.14	553.27	0.84	12.86	4.07	2.49
FantasyTalking (Wang et al., 2025)	14B	0.35	33.08	509.56	2.16	11.72	3.68	2.15
OmniAvatar-1.3B (Gan et al., 2025)	1.3B	0.97	**31.55**	**483.94**	3.16	10.91	3.70	2.18
Ours (OmniAvatar)	1.3B	**24.82**	33.96	486.77	**3.71**	**10.79**	**4.08**	**2.50**
CelebV-HQ								
Ground Truth	-	-	-	-	5.36	8.13	4.31	2.82
OmniAvatar-14B (teacher) (Gan et al., 2025)	14B	0.20	21.25	406.78	4.81	8.85	4.03	2.59
AniPortrait (Wei et al., 2024)	2.5B	0.38	23.87	427.43	1.14	12.07	4.24	**2.79**
FantasyTalking (Wang et al., 2025)	14B	0.35	26.87	454.98	2.48	10.95	3.93	2.46
OmniAvatar-1.3B (Gan et al., 2025)	1.3B	0.97	**22.28**	**382.35**	3.09	10.40	3.98	2.52
Ours (OmniAvatar)	1.3B	**24.82**	25.37	437.97	**3.78**	**10.08**	**4.29**	**2.79**

on; the MiniMax-H3 student is bidirectional like its teacher, so it needs no ODE initialization (2) and uses (1), (3), and the higher learning rates of (4).

Evaluation Protocol. We conduct evaluations under two settings: (1) **single-round evaluation**: standard audio-driven video generation with a single reference image and audio clip, evaluated on established benchmarks (HDTF, AVSpeech, CelebV-HQ); and (2) **transfer evaluation**: MiniMax-H3 is evaluated on 64 held-out conditions, using the Sync-C/D and IQA/ASE metrics of Tab. 1 on the generated video and its generated audio. We also build LiveTalk, a real-time interactive system with our distilled model as its renderer (Section 4.5), and evaluate it in multi-turn conversations (Appendix C).

4.2 Single-Round Evaluation

We evaluate our distilled model against several baselines: AniPortrait (Wei et al., 2024), FantasyTalking (Wang et al., 2025), and the bidirectional OmniAvatar-1.3B (Gan et al., 2025), whose trajectories initialize our student, and OmniAvatar-14B (Gan et al., 2025), our DMD teacher. The evaluation is conducted on 110 5-second clips from HDTF (Zhang et al., 2021) for in-domain evaluation, and on 100 randomly sampled 5-second clips from each of AVSpeech (Ephrat et al., 2018) and CelebV-HQ (Zhu et al., 2022) for out-of-domain evaluation. Following the evaluation protocol in Gan et al. (2025); Chen et al. (2025), we employ FID (Heusel et al., 2017), FVD (Unterthiner et al., 2019), IQA (Wu et al., 2023), and ASE (Wu et al., 2023) to assess visual quality and aesthetics, and Sync-C/D (Chung & Zisserman, 2017) to measure lip-sync synchronization between the audio condition and the generated lip movements. Ground truth videos are also evaluated on reference-free metrics for comparison. Throughput is measured at 512×512 resolution on a single GPU and averaged over multiple runs with randomly sampled conditions after warm-up generations.

Results and Analysis. As shown in Tab. 1, at the same model size our model improves lip-sync (Sync-C/D), IQA, and ASE over the bidirectional many-step variant OmniAvatar-1.3B (Gan et al., 2025) across all three benchmarks, while achieving 24.82 FPS (over 20× speedup). Its FID and FVD are higher than those of OmniAvatar-1.3B. FID and FVD measure how close the generated videos are to the real videos as a distribution and ignore the audio, while IQA and ASE, which score each video on its own, favor our model on all three benchmarks. On HDTF the ablation traces most of the FID gap to one step of the recipe: before the doubled learning rates and teacher CFG 6, the distilled model reaches FID 11.67, close to OmniAvatar-1.3B (10.85); this step raises Sync-C from 4.15 to 4.50 and FID to 13.68 (Tab. 3). Stronger guidance trades distribution coverage for fidelity to the condition, which raises FID (Ho & Salimans, 2022). In audio-driven talking video the audio is the

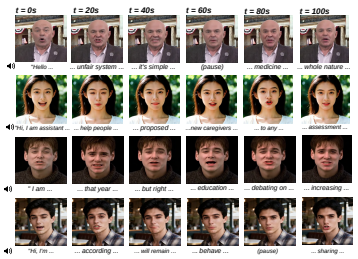

Figure 4: **Samples from our distilled OmniAvatar model.** Temporally coherent video with natural expressions, accurate lip-sync, and consistent identity.

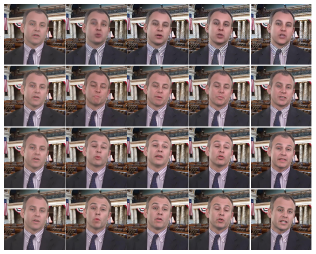

Figure 5: **Ablation examples.** Videos as each component of Tab. 3 is added: (1) curated conditions, (2) converged ODE initialization, (4) doubled learning rates with teacher CFG 6.

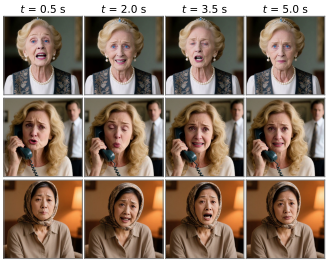

Figure 6: **Samples from our distilled MiniMax-H3 student** (four steps, DMD with the full-sample target, held-out conditions). Video and speech are generated jointly.

Table 2: **Distilling MiniMax-H3.** Bidirectional four-step LoRA student after 300 updates, 64 held-out clips. Full-sample target: the audio-energy target of Section 3.2, which matches the mean audio-latent energy of each generated clip to that of the teacher's own 49-step sample.

Model	**Sync-C** ↑	**Sync-D** ↓	**IQA** ↑	**ASE** ↑
Teacher (49 steps)	6.59	8.06	4.72	3.55
Student, DMD only	7.37	8.11	4.63	3.43
Student, DMD + full-sample target	**8.02**	**7.87**	4.59	3.36

condition that matters most, so we keep this step, and it produces visibly stronger mouth movements (Fig. 5). Figure 4 shows samples generated by our model.

4.3 Transfer to MiniMax-H3

On MiniMax-H3 (setup in Section 4.1) we compare DMD alone with DMD plus the full-sample target (Tab. 2). The target raises Sync-C from 7.37 to 8.02 and lowers Sync-D, while IQA and ASE change by at most 0.07. Higher learning rates, (4), also help here: raising the student learning rate from 2×10^{-5} to 5×10^{-5} increases Sync-C by 0.23. On one H200 the four-step student generates a five-second clip in 15.5 s against the teacher's 73.7 s (4.8×). Fig. 6 shows samples generated by the student.

4.4 Ablations

On OmniAvatar we ablate (1) curated multimodal conditions, (2) converged ODE initialization, and (4) the stronger learning signal, split into doubled learning rates and teacher score CFG 6 (Section 3.2). The full-sample target (3) is evaluated on MiniMax-H3 (Section 4.3). We evaluate on the same 110 HDTF clips as Section 4.2; each row of Tab. 3 adds one component.

Results and Analysis. The baseline uses 4000 conditions (2000 each from Hallo3 and HDTF) filtered by brightness and image quality, with the default Self Forcing settings: ODE initialization for 4000 steps, then 1000 DMD steps (critic and generator learning rates 4×10^{-7} and 2×10^{-6}, teacher score CFG 4, EMA decay 0.99); its videos degrade severely. Curated conditions remove the severe degradation, but blur and flicker remain; training the ODE initialization to convergence (20000 steps) removes them and gives the best FID, FVD, IQA, and ASE. Doubling both learning rates and raising the teacher score CFG to 6 improves lip-sync (Sync-C $4.15 \rightarrow 4.50$), with more pronounced mouth movements, and raises FID from 11.67 to 13.68 and FVD from 169.75 to 190.07. Applying all other improvements to the baseline conditions leaves the videos degraded (FID 23.89), with poor temporal consistency and abrupt color shifts between frames, so curated conditions are essential. Fig. 5 shows examples from four configurations.

Table 3: Ablation results with each improvement added sequentially. "Final Config" applies (1), (2), and (4) and is our benchmark model. The last row shows the Final Config w/o curated multimodal conditions (i.e., w/ baseline multimodal conditions).

Setting	FID↓	FVD↓	Sync-C↑	Sync-D↓	IQA↑	ASE↑
Baseline	27.10	338.08	3.13	11.77	3.95	2.38
+ Curated Multimodal Conditions	14.90	217.68	3.53	11.47	3.99	2.31
+ Converged ODE Initialization	**11.67**	**169.75**	4.15	11.19	**4.18**	**2.56**
+ Aggressive Learning Rate	12.10	179.73	4.29	11.07	4.15	2.53
+ Tuned Teacher Score CFG (**Final Config**)	13.68	190.07	**4.50**	**11.00**	4.13	2.52
Final Config **w/o** Curated Multimodal Conditions	23.89	261.19	3.85	11.42	4.06	2.47

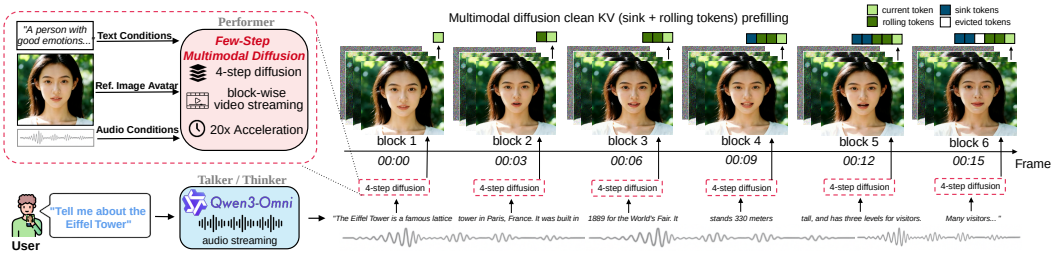

Figure 7: **Overview of the LiveTalk system.** Qwen3-Omni (Xu et al., 2025) answers the user's audio or text query with streaming speech; our four-step model renders the avatar from this audio, the reference image, and a text prompt, one block of 3 latent frames at a time.

4.5 Application: A Real-Time Interactive Avatar

The distilled model is fast enough to serve as the renderer of a conversational system. We build LiveTalk (Fig. 7), in which Qwen3-Omni (Xu et al., 2025) reasons over the user's audio or text and streams its spoken reply, and our four-step model renders the talking avatar from this audio, the reference image, and a text prompt, block by block. Three system choices keep the stream smooth. Audio is encoded in overlapping windows (1.75 s window, 0.75 s stride), so each block sees acoustic context across block boundaries without waiting for the full reply. Early high-fidelity frames stay in the KV cache as identity anchors, and the cache is refreshed periodically, which preserves the speaker's identity over minutes of video. Denoising of the current block runs in parallel with VAE decoding of the previous one, so generation stays ahead of playback. Appendix A gives the details, and Appendix C evaluates LiveTalk in multi-turn conversations.

5 Conclusion

We studied on-policy distillation of multimodal video diffusion models and found that the student's rollout loses what it should carry, at two levels. First, it loses the conditions: the image condition enters the rollout and, with it, the training loop, so curated conditions with a converged ODE initialization keep the conditions in the loop. Second, it loses the teacher's trajectory: after the ODE initialization the student sees only the DMD score of its own rollouts; a clip-level target from one full teacher sample per condition brings back a piece of it and improves lip-sync on MiniMax-H3, and higher learning rates with stronger teacher guidance make the student learn more from the teacher's score. The distilled four-step model runs at 24.82 FPS on a single GPU and improves lip-sync, image quality, and aesthetics over the bidirectional model of the same size. Beyond talking avatars, the same two questions, what the rollout starts from and how much of the teacher's signal reaches it, apply to any on-policy distillation whose conditions enter the generated output.

Ethics Statement

Our model generates realistic talking-head videos from a single image and an audio clip, which could be misused. Generated videos should be clearly labeled as synthetic. The conversation demo in the supplementary material animates a person for research illustration only: the avatar's speech and video are both generated by our system, and the user's face is pixelated.

Reproducibility Statement

Section 4.1 describes the models, data, and evaluation protocol. Tabs. 6 and 7 list the training hyperparameters, and Appendices D and D.3 describe the evaluation criteria and settings. Video samples are in the supplementary material (Appendix G).

AI Use Statement

We used LLMs to polish parts of the writing and to help write code for later analyses. As described in Section 3.2, Qwen-Image and Qwen2.5-VL generate part of the training conditions. The authors verified all AI-assisted content and take full responsibility for the paper.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. TRACT: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

BrianChen1129. Question about Image2Video distillation implementation. GitHub issue, 2025. URL https://github.com/tianweiy/CausVid/issues/51. Issue #51, tianweiy/CausVid repository.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

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.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. HunyuanVideo-Avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, volume 10117 of *Lecture Notes in Computer Science*, pp. 251–263. Springer, 03 2017. ISBN 978-3-319-54426-7. doi: 10.1007/978-3-319-54427-4_19.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

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. *arXiv preprint arXiv:1804.03619*, 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.

Yu Gao, Haoyuan Guo, Tuyen Hoang, et al. Seedance 1.0: Exploring the boundaries of video generation models, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*, pp. 1593–1597, 2023.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling distribution matching for flow-based text-to-image distillation, 2025. URL https://arxiv.org/abs/2506.00523.

Google DeepMind. Veo 3 technical report. 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. BOOT: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025a.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live Avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. FullDiT: Multi-task video generative foundation model with full attention. *arXiv preprint arXiv:2503.19907*, 2025.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, et al. Avatar Forcing: Real-time interactive head avatar generation for natural conversation, 2026.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the association for computational linguistics ACL 2024*, pp. 11286–11315, 2024.

Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. HalloLive: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. *arXiv preprint arXiv:2604.23632*, 2026a.

Haoyu Li, Tingyan Wen, Lin Qi, Zhe Wu, Yihuang Chen, Xing Zhou, Lifei Zhu, Xueqian Wang, and Kai Zhang. 1.x-Distill: Breaking the diversity, quality, and efficiency barrier in distribution matching distillation. *arXiv preprint arXiv:2604.04018*, 2026b.

Mingxiao Li, Bo Wan, Marie-Francine Moens, and Tinne Tuytelaars. Animate your motion: Turning still images into dynamic videos. In *European Conference on Computer Vision*, pp. 409–425. Springer, 2024.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. STIV: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16249–16259, 2025.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025b.

Xiangyu Liu, Feng Gao, Xiaomei Zhang, Yong Zhang, Xiaoming Wei, Zhen Lei, and Xiangyu Zhu. TurboTalk: Progressive distillation for one-step audio-driven talking avatar generation. *arXiv preprint arXiv:2604.14580*, 2026.

Chetwin Low and Weimin Wang. TalkingMachines: Real-time audio-driven FaceTime-style video via autoregressive diffusion models. *arXiv preprint arXiv:2506.03099*, 2025.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Rang Meng, Weipeng Wu, Yuming Li, and Chenguang Ma. EchoTorrent: Towards swift, sustained, and streaming multi-modal video generation. *arXiv preprint arXiv:2602.13669*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. Blog post, `https://www.minimax.io/blog/minimax-h3`; model card `https://huggingface.co/MiniMaxAI/MiniMax-H3`, 2026. Released 2026-07-31.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Qwen Team. Qwen3-VL-30B-A3B-Instruct. Model card, https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. *arXiv preprint arXiv:1910.00292*, 2019.

Le Shen, Qian Qiao, Tan Yu, et al. SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Yaofeng Su, Yuming Li, Zeyue Xue, et al. OmniForcing: Unleashing real-time joint audio-visual generation, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. URL https://arxiv.org/abs/1812.01717.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9891–9900, 2025.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels, 2023. URL https://arxiv.org/abs/2312.17090.

Zhuguanyu Wu, Ruihao Gong, Yang Yong, Yushi Huang, Xiangyu Fan, Lei Yang, Dahua Lin, and Xianglong Liu. SGMD: Score gradient matching distillation for few-step video diffusion distillation. *arXiv preprint arXiv:2605.30116*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Xilluill. Question about ode_init.pt weight and DMD training. GitHub Issue, 2025. URL https://github.com/guandeh17/Self-Forcing/issues/50. Issue #50, Self-Forcing repository.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Jintao Zhang, Kai Jiang, Jintao Chen, et al. Vidu S2: Real-time interactive, editable, and spatial video generation, 2026.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3661–3670, 2021.

Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, and Zhou Zhao. FADA: Fast diffusion avatar synthesis with mixed-supervised Multi-CFG distillation, 2024.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

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. *arXiv preprint arXiv:2602.02214*, 2026.

A LiveTalk System Details

Based on the distilled model, we build LiveTalk (Fig. 7), a complete real-time multimodal interactive avatar system integrating our model with Qwen3-Omni (Xu et al., 2025) for end-to-end visual communication. Our modular architecture comprises two key components: a real-time performer module (our distilled video diffusion model) that renders synchronized talking avatars, and a thinker/talker module (Qwen3-Omni) that handles reasoning and generates streaming audio responses. We detail the system pipeline overview, and the three critical system-level challenges below.

System Pipeline Overview. Figure 7 illustrates LiveTalk's interaction pipeline. When a user provides audio or text input, Qwen3-Omni (Xu et al., 2025) generates streaming audio responses. Our video diffusion model takes three multimodal conditions: (1) streaming audio output c_{audio} from Qwen3-Omni, (2) reference avatar image c_{img} defining visual identity, and (3) text prompts c_{text} describing the desired motion (e.g., "A person speaking naturally with expressive gestures and emotions"). These jointly drive block-wise autoregressive generation, where each block of $b = 3$ latent frames undergoes $k = 4$ diffusion steps. Clean KV cache from previous blocks is prefilled for visual consistency, enabling streaming synchronized to audio.

Streaming Audio Conditioning. Audio-driven video generation requires acoustic context from adjacent frames for smooth lip-sync and natural motion. Conditioning each video block solely on temporally aligned audio causes discontinuities at block boundaries. Waiting for extended audio sequences (encoding entire clips before generation) introduces prohibitive latency. Our solution

uses overlapped windowing: we encode and generate as soon as a small windowed segment becomes available. Specifically, we use a 1.75s audio window with a 0.75s stride, where 1.0s of contextual audio from the previous window provides sufficient overlap to ensure smooth transitions across block boundaries.

Long-Horizon Video Streaming with Speaker Identity Preservation. Interactive dialogues often last for minutes, requiring the generated avatar's identity to remain consistent over long time spans. Although Self Forcing (Huang et al., 2025a) reduces train–test mismatch by aligning training and inference distributions, identity still degrades beyond the training window as accumulated errors cause color drift and geometric distortions. To counteract this drift, we use a training-free method, **Anchor-Heavy Identity Sinks (AHIS)**: we allocate part of the KV cache as attention sinks (Xiao et al., 2024) that permanently store early high-fidelity speaker frames as identity anchors. With a 5-latent-frame attention window and 3-frame blocks, the steady-state cache holds these anchor frames together with the current block. For minute-scale generation we additionally refresh the KV cache periodically at inference time; the identity anchors combined with this periodic recaching preserve the speaker's identity over minutes of generated video.

Parallel Pipeline for Video Denoising and Decoding. Streaming video generation requires diffusion denoising (predicting clean latents) and VAE decoding (converting to pixels). Sequential execution risks playback stalling when generation time exceeds playback duration. We adopt pipeline parallelism: while the current block undergoes denoising, the previous block is simultaneously decoded. This reduces the interval between decoded blocks from the sum to the maximum of the two stage durations, ensuring generation stays ahead of playback and enabling non-stalling streaming with seamless real-time rendering.

B Extra Quantitative Results on Single-Round Evaluation

We extend the single-round evaluation of Section 4.2 to 500 clips per dataset. Tab. 4 confirms the result at this scale: our distilled model improves lip-sync (Sync-C/D), IQA, and ASE over the bidirectional many-step variant OmniAvatar-1.3B (Gan et al., 2025) across all three benchmarks.

Table 4: Quantitative results on the extended evaluation sets (500 clips per dataset). OmniAvatar-14B (grey) is the teacher and is not ranked. **Bold**: best; underline: second best.

Methods	FID↓	FVD↓	Sync-C↑	Sync-D↓	IQA↑	ASE↑
HDTF						
OmniAvatar-14B (teacher)	5.94	64.86	6.24	8.80	4.31	2.72
AniPortrait (2.5B)	10.88	127.52	1.09	13.28	4.48	2.86
FantasyTalking (14B)	11.77	216.29	2.89	11.73	4.10	2.55
OmniAvatar-1.3B	**6.72**	**81.72**	3.89	10.92	4.26	2.67
Ours (OmniAvatar-1.3B)	9.19	90.17	**4.51**	**10.61**	**4.51**	**2.88**
AVSpeech						
OmniAvatar-14B (teacher)	11.53	134.33	5.29	9.08	4.06	2.39
AniPortrait (2.5B)	14.15	180.46	0.93	13.01	4.32	2.58
FantasyTalking (14B)	13.27	187.32	2.30	11.57	3.96	2.31
OmniAvatar-1.3B	**12.05**	**138.04**	3.31	10.74	3.99	2.32
Ours (OmniAvatar-1.3B)	14.31	166.29	**3.83**	**10.43**	**4.37**	**2.63**
CelebV-HQ						
OmniAvatar-14B (teacher)	10.05	129.36	4.65	8.75	4.20	2.72
AniPortrait (2.5B)	11.99	148.76	0.85	12.16	4.42	**2.91**
FantasyTalking (14B)	13.00	174.05	2.25	10.91	4.05	2.58
OmniAvatar-1.3B	**10.36**	**124.73**	2.98	10.27	4.13	2.63
Ours (OmniAvatar-1.3B)	13.03	144.86	**3.51**	**10.08**	**4.44**	2.89

C LiveTalk Multi-Round Interaction Evaluation

Existing benchmarks for multimodal-driven avatar video generation primarily focus on single-audio-clip metrics, using traditional visual evaluation measures such as lip-sync accuracy (Sync-C/D), image quality (FID/FVD/IQA), and aesthetics (ASE). However, these metrics fail to assess multi-turn interaction quality, critical for real-world multimodal conversational applications. To address this gap, we propose a multi-round interaction benchmark and an evaluation protocol based on

vision-language models (VLMs) (Bai et al., 2023; Lee et al., 2024) to comprehensively evaluate the multi-turn conversational capabilities of audio-driven avatar systems.

Benchmark Design Methodology. We curate 100 multi-turn evaluation scenarios that require multimodal interactive AI systems to provide coherent audio-driven video generation responses across conversational turns. For example, in the first round, a user might ask, "Tell me about the Eiffel Tower," and the system should generate a coherent introduction to the landmark with synchronized audio and video. In the subsequent round, the user might follow up with, "Where is this avenue located?" referencing the Champs-Élysées mentioned in the previous response. The AI system must generate coherent video and audio outputs that are contextually grounded in its previous multimodal conversational history, demonstrating the ability to maintain temporal coherence and contextual awareness across multiple interaction turns.

Evaluation Protocol. We adopt Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025) as the VLM-as-evaluator (Lee et al., 2024) with structured prompts tailored to each dimension. Our evaluation framework comprises two primary categories: *Visual Interaction Performance* (4 dimensions: Emotional Appropriateness, Nonverbal Interaction, Multi-Video Coherence, and Conversational Naturalness) and *Interaction Content Quality* (5 dimensions: Semantic Relevance, Information Completeness, Logical Consistency, Context Understanding, and Overall Interaction Experience). Appendix D describes each dimension. System-generated audio responses are transcribed via FunASR (Gao et al., 2023) for content quality evaluation.

We normalize raw scores for each metric using z-score transformation. Specifically, for each metric, we create a single pooled distribution containing scores from all methods on all evaluation samples, compute z-scores as $z = (x - \mu)/\sigma$ where μ and σ are from the pooled distribution, and convert each z-score to its percentile (0-100). We report average percentiles per baseline for each metric, providing relative performance interpretation.

Results and Analysis. Table 5 presents multi-round evaluation against Veo3 (Google DeepMind, 2025) (API model `veo-3.0-fast`). LiveTalk outperforms Veo3 on all nine dimensions. Notably, Veo3 exhibits visual drift across turns, as it lacks a memory mechanism for multi-round video interaction, and its generation time (61 s per turn) breaks the conversational flow. In contrast, LiveTalk leverages AR generation with KV cache to maintain visual memory states and the Qwen3-Omni thinker/talker module to preserve textual memory states, enabling coherent real-time multimodal generation.

Table 5: Multi-Round Interaction Evaluation. We benchmark LiveTalk against Veo3 on the proposed interaction benchmark. We report Z-Score percentile values for Visual Interaction Performance (VIP) and Interaction Content Quality (ICQ) metrics, along with average inference throughput (Thr.) and latency (Lat.) per turn. **Bold** indicates best performance. *Note:* EA: Emotional Appropriateness, NI: Nonverbal Interaction, MVC: Multi-Video Coherence, CN: Conversational Naturalness; SR: Semantic Relevance, IC: Information Completeness, LC: Logical Consistency, CU: Context Understanding, OIE: Overall Interaction Experience.

Method	Thr. (FPS) ↑	Lat. (s) ↓	VIP ↑				ICQ ↑				
			EA	NI	MVC	CN	SR	IC	LC	CU	OIE
Veo3	-	61.46	31.26	31.95	19.04	32.84	29.95	20.24	25.62	25.83	22.29
LiveTalk	24.82	1.16	**71.12**	**68.91**	**79.75**	**66.33**	**76.20**	**78.78**	**78.45**	**77.81**	**79.60**

D Multi-Round Evaluation Details

This appendix provides a comprehensive description of the nine evaluation dimensions comprising our interaction benchmark. Each dimension is assessed by a VLM using structured prompts that define specific scoring criteria. The evaluation framework is divided into two major categories: **Visual Interaction Performance** and **Interaction Content Quality**. We also provide extra implementation details for the evaluation protocol.

D.1 Visual Interaction Performance

Visual interaction performance captures the non-verbal and emotional aspects of the assistant's video responses, emphasizing the quality of human-like engagement and visual consistency across conversational turns.

Emotional Appropriateness. This dimension evaluates the degree to which the assistant's facial expressions align with conversational context and emotional content.

Assessment criteria include:

- **Content-Emotion Matching:** Appropriateness of emotional expressions relative to semantic content (e.g., conveying seriousness during complex explanations; displaying warmth during positive encouragement).
- **Intensity Calibration:** Proportionality of emotional intensity to the significance and tone of the conversational content.
- **Emotional Transitions:** Naturalness and coherence of emotional shifts across multiple consecutive video responses.
- **User Responsiveness:** Adaptive emotional responses that acknowledge and appropriately react to the user's inferred emotional state (e.g., providing reassurance for confusion, matching enthusiasm).

Nonverbal Interaction. This dimension quantifies the quality of interactive nonverbal cues that establish and maintain user engagement.

Assessment criteria include:

- **Eye Contact Quality:** Maintenance of natural, engaging gaze patterns that simulate direct interaction with the user.
- **Interactive Gestures & Micro-expressions:** Effective use of confirmatory nods, thoughtful pauses, raised eyebrows, and other subtle movements that signal active participation in dialogue.
- **Listening Response:** Presence of nonverbal acknowledgment signals that indicate comprehension and processing of user input prior to verbal response.
- **Interaction Authenticity:** Overall impression of genuine bidirectional dialogue rather than unidirectional content delivery.

Multi-Video Coherence. This dimension assesses visual and behavioral consistency across multiple video responses within the conversation.

Assessment criteria include:

- **Visual Identity Consistency:** Stability of the speaker's physical appearance, including facial features, hairstyle, and attire.
- **Scene Consistency:** Stability of environmental factors including background elements, lighting conditions, and camera perspective.
- **Emotional Continuity:** Logical progression and smooth evolution of emotional states from one video response to subsequent responses.
- **Behavioral Coherence:** Consistency in posture, gestural patterns, speech cadence, and overall presentation style.

Conversational Naturalness. This dimension measures the degree to which the interaction conveys authentic human-like spontaneity rather than scripted or mechanistic behavior.

Assessment criteria include:

- **Dialogue vs. Broadcast Distinction:** Presence of interactive, bidirectional conversational dynamics versus formal, unidirectional presentation style.
- **Spontaneous Human Characteristics:** Natural occurrences of brief pauses, thinking moments, speech hesitations, and other markers of authentic spontaneous communication.

D.2 Interaction Content Quality

Interaction content quality evaluates the semantic, logical, and contextual appropriateness of the assistant's verbal responses.

Semantic Relevance. This dimension assesses whether the assistant's responses directly and accurately address the topical focus of user queries.

Information Completeness. This dimension evaluates if the provided answers are comprehensive, sufficiently detailed, and self-contained.

Logical Consistency. This dimension identifies contradictions, factual errors, or logical inconsistencies both within individual responses and across the entire conversational history.

Context Understanding. This dimension measures the assistant's ability to maintain conversational state, correctly resolve anaphoric references, and effectively build upon information established in prior dialogue turns.

Overall Interaction Experience. This dimension provides a holistic assessment encompassing conversational fluency, effectiveness in task completion, and overall user satisfaction.

D.3 Evaluation Protocol Details

VLM Evaluator. We employ Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025) as our evaluation model, deployed via vLLM (Kwon et al., 2023). The model configuration supports contexts up to 256K tokens and processes up to 32 frames per video at 784×784 resolution. We use a sampling temperature of 0.1 combined with top-p sampling (p=0.9). Each evaluation dimension is guided by a detailed rubric encoded in dimension-specific prompt files, with model responses structured as `\boxed{score}` for automated parsing and aggregation.

ASR Transcription. Full conversational transcripts are constructed by transcribing both user audio inputs (16kHz WAV/MP3 format) and assistant video audio tracks into English text. We utilize the FunASR (`speech_paraformer-asr-en-16k-vocab4199`) (Gao et al., 2023) model, which employs a non-autoregressive Paraformer architecture.

Multi-Video Evaluation. For conversations containing multiple assistant video responses, all video files are presented to the VLM evaluator in chronological order alongside the complete conversational transcript. This multi-video presentation enables both isolated analysis of individual video segments and comparative cross-video analysis, which is essential for accurately assessing the *Multi-Video Coherence* dimension and evaluating *Emotional Transitions*. All content quality dimensions leverage the full conversation history to assess semantic coherence, logical consistency, and contextual understanding across all dialogue turns.

E Training Dynamics of DMD under Multimodal Conditioning: Extended Analysis

While instability of DMD-based distillation has been reported by the community (Ge et al., 2025; BrianChen1129, 2025), a gradient-level account of where the instability enters under multimodal conditioning has been missing. DMD training has three parts that depend on each other in a cycle: the generator g_ϕ produces rollouts, the critic s_ψ learns to denoise them, and the gradient $s_\psi - s_\theta$ (where s_θ is the frozen teacher) drives the generator toward the teacher's distribution. Any degradation in one component cascades to the others. We analyze how three of our improvements, curated conditions, a converged ODE initialization, and a tuned teacher CFG, each address a distinct failure mode in this loop, using five configurations over the first 800 DMD steps.

E.1 Converged ODE Initialization: A Better Starting Point for the Critic

The critic s_ψ trains exclusively on the generator's outputs. When the generator starts from a weak ODE checkpoint, the critic must denoise low-quality samples—a harder task, reflected in a higher critic loss.

Comparing *Baseline* (ODE 4k) with +*Converge* (ODE 15k, with doubled learning rates and teacher CFG 6), Fig. 3e shows the combined effect of these changes: critic loss decreases by 9%, DMD gradient norm rises by 24.5%, and generator loss rises by 61.6%. In Fig. 3a, the +*Converge* curve consistently runs below *Baseline*, with the largest gap in the [100–400) step range. The combined

changes lower the critic loss and strengthen the DMD gradients from the start of DMD training (Fig. 3a,b).

However, without curated conditions a converged ODE initialization is insufficient. Fig. 3c shows that +*Converge* (orange) exhibits the **most unstable** generator gradient norm, with prominent spikes at steps 300–500. The baseline conditions keep training unstable, motivating the next improvement.

E.2 Curated Multimodal Conditions: Stabilizing the Gradients

Low-quality reference images and vague text prompts corrupt **both sides** of the DMD gradient: the critic trains on noisy rollouts while the teacher evaluates under noisy conditions, so both terms of the gradient $s_\psi - s_\theta$ carry these flaws.

The configuration with curated conditions is the most stable. In Fig. 3a, +*Curated Cond*+*Converge* (green) is both lower and visibly smoother than +*Converge* (orange). Fig. 3d quantifies this: the standard deviation of the critic loss over steps 100–800 falls from 0.105 to 0.085. The effect is most striking in Fig. 3c, where the violent spikes of +*Converge* are almost entirely suppressed. Fig. 3d confirms that curated conditions achieve the lowest DMD gradient norm standard deviation across all configurations.

These reductions in **variance**—not just mean—are the key signature of stable optimization. The ablation study (Tab. 3) corroborates this: applying all other improvements **without** curated conditions yields FID=23.89 against 27.10 for *Baseline*, whereas curated conditions alone reduce FID to 14.90. Condition quality is therefore the foundation of the recipe.

E.3 Tuned Teacher CFG: Strengthening the Conditional Signal

With the default teacher CFG (γ=4), the conditional signal is too weak for the generator to learn strong lip-sync.

Fig. 3f–h compares three CFG scales. Increasing CFG to 6 raises the DMD gradient norm by 12% (panel g), while keeping the generator loss stable (panel f) and **reducing** the generator gradient norm (panel h). Together with the doubled learning rates, this improves lip-sync (Sync-C: $4.15 \rightarrow 4.50$).

We further explored a higher CFG scale of 8, but observed that the benefits did not continue. Instead, CFG=8 destabilizes training (Fig. 3f–h, purple dashed). The generator loss becomes highly unstable with substantially increased variance (panel f), and the DMD gradient norm rises by a further 12% (panel g). Meanwhile, the critic loss remains nearly unchanged between CFG=6 and CFG=8 (mean over steps 0–800: 0.105 versus 0.107), while the teacher signal is amplified, $s_{\theta,\text{cfg}} = s_{\theta,\text{uncond}} + \gamma(s_{\theta,\text{cond}} - s_{\theta,\text{uncond}})$, which dominates the DMD gradient and drives the generator into an unnatural training regime. Under multimodal conditioning, the dominance of the CFG-augmented teacher term in the DMD objective (Liu et al., 2025a) thus destabilizes the generator once the scale is too high.

E.4 Synergy: How the Three Improvements Build on Each Other

The three remedies used on OmniAvatar target complementary levels of the optimization (Fig. 3d) and build on each other:

- **(1) Curated conditions** establish a clean signal floor. Without them, neither ODE convergence nor CFG tuning compensates (Tab. 3: FID=23.89 without curated conditions vs. 13.68 with curated conditions).
- **(2) Converged ODE** builds on this floor with a strong starting point, giving the critic better samples from step 1. Without clean conditions, it amplifies noise instead (Fig. 3c, orange spikes).
- **(4) Doubled learning rates and teacher CFG 6** exploit the resulting stable dynamics to strengthen learning from the teacher's score.

This gives an ordering principle: **first ensure clean inputs, then strong initialization, and only then amplify the learning signal**.

F Training Configuration Details

For OmniAvatar, the distillation pipeline consists of two sequential stages: ODE initialization and on-policy distribution matching distillation (DMD) (Section 3.1). Tab. 6 details the hyperparameters and optimization settings for both stages. For MiniMax-H3, Tab. 7 lists the DMD settings; there is no ODE initialization.

Table 6: **Training configurations for ODE trajectory initialization and DMD distillation.** CFG (Generating Trajectory) denotes the classifier-free guidance scale used by the bidirectional model (OmniAvatar-1.3B) during ODE trajectory generation. For DMD distillation, the critic score network s_ψ is first updated for 20 warm-up steps, after which the student generator network g_ϕ is trained (Section 3.1). Exponential moving average (EMA) updates are enabled from step 200 onward.

Hyperparameters	**ODE Initialization**
Batch size	64
Learning rate	$4\text{e}{-5}$
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}{-8}$, weight_decay=0
CFG (Generating Trajectory)	4.5
Hyperparameters	**DMD Distillation**
Teacher score network	OmniAvatar-14B
Teacher score CFG	6.0
Critic score network	OmniAvatar-1.3B
Batch size	64
Optimizer (g_ϕ)	AdamW, $\beta_1 = 0$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}{-8}$, weight_decay=0.01
Optimizer (s_ψ)	AdamW, $\beta_1 = 0$, $\beta_2 = 0.999$, $\epsilon = 1\text{e}{-8}$, weight_decay=0.01
Learning rate (g_ϕ)	$4\text{e}{-6}$
Learning rate (s_ψ)	$8\text{e}{-7}$
Critic updates per generator update	5, with critic warmup for 20 steps
EMA decay	0.99

Table 7: Training configuration for distilling MiniMax-H3 (the DMD-only run omits the auxiliary target).

Student / critic	Low-rank adapters, rank 64 / 128, on attention and feed-forward projections of all 50 blocks
Teacher score network	MiniMax-H3 (frozen), guidance built in, one query per score
Optimizer (both)	AdamW, β_1=0.9, β_2=0.999, weight_decay=0.01
Learning rate (student / critic)	5×10^{-5} / 2×10^{-5}, 20-update linear warm-up
Gradient clipping	1.0
Batch size	8 (1 per GPU, 8 GPUs)
Updates	300 student updates, 3 critic updates per student update
Sampling timesteps	999, 749, 500, 250; shift 12 (video), 3 (audio)
Training conditions	240 curated HDTF conditions, 512×512, 124 frames
Auxiliary target	Weight 0.1, ramped over 30 updates; reference: audio-latent energy of the teacher's 49-step sample
Evaluated weights	Student after update 300 (no EMA)

G Video Demos and Qualitative Examples

Video examples are provided in the supplementary material: eight super-resolved clips from our distilled OmniAvatar model on held-out conditions; four MiniMax-H3 conditions, each shown with the teacher, the DMD-only student, and the student trained with the full-sample target (Tab. 2), and four further samples of the latter student (Fig. 6); and a real-time conversation with the LiveTalk system. Fig. 2 shows the collapse of Section 3.2. Fig. 4 shows frames generated by our distilled

OmniAvatar model, and Fig. 5 shows the same conditions under four configurations of the ablation in Section 4.4.