MEASURE BEFORE YOU OPERATE: LAYER-WISE GRADIENT SURGERY IN AUDIO–VIDEO DIFFUSION RL IS NECESSARY ON SOME BACKBONES AND INERT ON OTHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

OmniNFT, the current recipe for RL post-training of joint audio–video diffusion models, protects the audio tower from video-loss gradients with a per-block scale ("layer-wise gradient surgery") whose constants were tuned on one backbone, LTX-2, and published as block indices. We ask three questions: where the relevant blocks are on other backbones, whether the surgery is needed at all, and whether its exact placement matters. Two probes that take hours rather than training runs locate the synchronization-critical blocks and measure per-block gradient leakage. On five architectures the zones sit at different relative depths, and no index-copying or rescaling heuristic recovers them; on LTX-2 itself our probes do not recover the published window. Within a model line the zones do not move under SFT, weight souping or RL. Sixteen controlled RL arms on three backbones then give different answers to the other two questions. Whether surgery is needed depends on the backbone: on a 29B model, fifty epochs of RL without it leave word error rate 39% above the SFT start while the training reward keeps rising, and the calibrated schedule from the same start lowers WER by 25% (p=0.002); on a 3B model and on the public Ovi, none of the schedules, including no surgery, can be told apart. Where the surgery is placed cannot be resolved at this budget on any backbone: differences between schedules are within noise and change sign between prompt sets. A first-order argument shows that the surgery acts through the product of leakage and gradient conflict, which explains the inert cases but not the 3B one. The recipe also contains the first speech-intelligibility reward for joint AV RL, which lowers WER by 28% on public Harvard sentences and 48% on the VABench speech subset, with smaller gains under independent ASR judges. Probe code and the public evaluation protocol will be released.

1 INTRODUCTION

RL with verifiable rewards is now a standard post-training step for generative models, and it has reached joint audio–video (AV) generation, where a single diffusion policy has to satisfy per-modality quality, cross-modal consistency and fine-grained synchronization at the same time. The strongest published recipe, OmniNFT (Zhang et al., 2026), adds three mechanisms on top of DiffusionNFT (Zheng et al., 2025): modality-wise advantage routing, region-wise loss reweighting, and layer-wise gradient surgery, a per-block scale $\alpha(b)$ on the gradient that the video loss sends into the audio tower through the audio→video cross-attention keys and values. The published schedule (α=0.1 on blocks 1–10, 0.3 on blocks 40–47 of LTX-2's 48) comes from ablations on LTX-2. The recipe does not say how to set it for another model, and the two obvious guesses, copying the indices or rescaling them with depth, both assume that the relevant blocks sit at fixed relative positions across architectures.

We test that assumption and, more importantly, ask whether the constants matter at all. We port the recipe to a 29B twin-tower AV transformer with 60 blocks, a 3B sibling with 32, and the public 11B Ovi (Low et al., 2025). On each we run two probes that take hours: muting one cross-modal attention direction over block ranges during rollout, which locates the synchronization-critical zone,

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.

and running audio-only and video-only backward passes, which measure how much video gradient reaches audio parameters in each block and whether it conflicts with the audio objective. We then train 16 RL arms across the three backbones that differ only in the surgery schedule, including no surgery.

The results split into three parts. First, the zones do not transfer but are stable: the synchronization zone lies at 83–98% of depth on the 29B model, 69–81% on the 3B and 33–48% on Ovi; index copying misses it wherever it is defined and proportional rescaling misses on two of three; the zone does not move under SFT, weight souping or RL, so one measurement per model line is enough. Second, whether surgery is needed depends on the backbone. On the 29B model RL without surgery never improves speech: from the SFT base, WER is 39–46% above the base at every checkpoint we evaluated, including the one the training reward selects, while the calibrated schedule from the same base lowers WER by 25% ($0.235 \rightarrow 0.175$, p=0.002). On Ovi and the 3B, seven and three schedules respectively, including no surgery, are indistinguishable. The audio-tower leak ratio separates the two extremes (0.16 vs. 0.014) but does not explain the 3B case (0.25, no effect). Third, fine placement is not resolvable at a 50-epoch budget: WER differences between surgery schedules are 0.01–0.02, flip between checkpoints and reverse between VABench and Harvard. A first-order analysis (Sec. 3) says why: the surgery acts through the product of leakage and conflict, which is small wherever we detect nothing.

What this means in practice. Measure the audio-tower leak ratio (two GPU-hours). If it is two orders of magnitude below one, as on Ovi, skip the surgery and do not spend an ablation on it. If it is around 0.1–0.3, apply it with the zones from probe 1: in our data this is never worse than no surgery and on the 29B backbone it is the only configuration that improves speech; do not spend compute on refining the placement. In either case, do not pick checkpoints by total reward alone; on two backbones the reward-selected checkpoint is worse in speech than a later one. See also Appendix, Fig. 3.

Separately, the recipe carries the first speech-intelligibility objective for joint AV RL: Whisper-transcribed WER against the scripted text, routed to the audio branch only. On the 29B backbone it cuts WER by 48% on the VABench speech subset and 28% on 100 public-domain Harvard sentences, with smaller gains under two independent ASR judges; the same objective works on the 3B model and on Ovi.

Contributions. (1) A two-probe protocol for cross-modal sensitivity and gradient leakage, applied to five architectures (three with RL ablations), with per-video statistics. (2) A controlled study of surgery transfer showing that the decision that matters, whether to apply the surgery, is a backbone property that the probes make cheap to take, while the decision that does not, where exactly to place it, is expensive and at a 50-epoch budget unnecessary; with a first-order account of when the mechanism can act. (3) A practical AV-RL recipe with a speech reward, validated on a public protocol and independent judges, plus a checklist of silent failure modes (Appendix K).

2 Related Work

RL post-training for diffusion. Policy-gradient fine-tuning of diffusion models began with PPO-style objectives on the denoising chain (Black et al., 2024; Fan et al., 2023) and differentiable rewards (Clark et al., 2024; Xu et al., 2023); group-relative optimization (Guo et al., 2025) was adapted through ODE-to-SDE conversions (Xue et al., 2025; Liu et al., 2025a; Li et al., 2025; He et al., 2025) and through forward-process optimization that keeps the deterministic sampler (Zheng et al., 2025), which we build on. Multi-reward normalisation (Liu et al., 2026b) addresses conflicting objectives within a modality; the modality-wise setting adds routing across branches and, with it, the leakage problem studied here. Several of our checklist items are instances of reward over-optimisation (Uehara et al., 2024).

Gradient conflict. Layer-wise surgery is a special case of gradient conflict in multi-task learning (Yu et al., 2020; Chen et al., 2018; Liu et al., 2021). Two things make the AV case simpler: the foreign gradient enters through a single, architecturally identified path (the cross-modal KV projections), and the two objectives live on disjoint parameter sets, so the interaction can be measured per block offline, once. Proposition 1 is the PCGrad first-order argument for this structure.

Joint audio–video generation and speech rewards. Most open joint AV generators are diffusion transformers with two modality streams coupled by cross-modal attention (HaCohen et al., 2026; Liu et al., 2025c; 2026a; Low et al., 2025; Wang et al., 2025); single-stream designs also exist. OmniNFT (Zhang et al., 2026) is the first RL recipe for this setting; we study how it transfers rather than propose a competitor. We keep its reward suite (Ma et al., 2025; Liu et al., 2025b; Tjandra et al., 2025; Wu et al., 2023; Iashin et al., 2024) and add lip-region synchronization (Li et al., 2024; Chung & Zisserman, 2016) and ASR scoring (Radford et al., 2023). ASR rewards are common in speech synthesis and talking-head generation, where the generator controls one modality; we are not aware of a joint AV-RL stack that optimises speech intelligibility. A concurrent technical report (Anonymous, 2026) documents the production training of our primary backbone; the measurement protocol, the cross-backbone evidence and the controlled ablations are new here.

3 Method: Probes, Schedule, and a First-Order Account

Setting. The generators we study process video and audio token streams through N paired blocks; each block has per-modality self-attention, text cross-attention and feed-forward layers, plus two cross-modal attentions, a2v (video queries, audio keys/values) and v2a (the reverse). Following Zheng et al. (2025); Zhang et al. (2026), the policy is trained on the forward process from deterministic ODE rollouts (20 steps, no CFG, K rollouts per prompt), with per-reward group-normalised advantages routed by modality (video rewards to the video branch, audio to the audio branch, synchronization to both), an anchor MSE to the frozen SFT prediction ($\beta{=}10^{-4}$), and training timesteps from the noisy 40% of the schedule. Layer-wise surgery multiplies, in block b, the gradient flowing through the a2v key/value input by $\alpha(b) \in [0, 1]$ via a straight-through hook; the forward pass is untouched. Region-wise reweighting is kept as published. Hyperparameters: Appendix B.

Probe 1: block-range KV-ablation. During a full RL-regime rollout we zero the output of one cross-modal direction over a contiguous block range (a forward hook on the value projection removes exactly that contribution) and generate a fixed 32-prompt set with one seed. The per-video change in Synchformer desynchronization relative to the unablated baseline (paired Wilcoxon over the 32 prompts) localises where the muted direction carries synchronization-relevant signal. A full decade sweep of one direction costs 15–20 GPU-hours.

Probe 2: gradient leakage and conflict. We replay training-style forwards on rollout latents at noisy timesteps and run two separate backward passes — a video-only and an audio-only reconstruction loss — recording per-block L2 norms of the audio-branch adapter gradients (audio tower and a2v-KV groups). The leak ratio $\rho(b) = \|\nabla_{\theta_a(b)} L_v\| / \|\nabla_{\theta_a(b)} L_a\|$ measures how strongly video supervision perturbs audio parameters at block b; where both towers fit on one device we also record the cosine $c(b) = \cos(\nabla_{\theta_a(b)} L_a, \nabla_{\theta_a(b)} L_v)$. The probe costs under two GPU-hours.

From probes to a schedule. The published schedule is reproduced with measured zones: $\alpha{=}0$ at block 0, 0.1 over the blocks where ρ is within a factor two of its peak (the leakage zone), 0.3 over the probe-1 synchronization zone, and 1 elsewhere. We call this Measured. Two closed forms follow from the same data and are also trained as arms: $\alpha(b) = \min(1, \tau/\rho(b))$ (bound the leaked gradient to a fraction τ of the native one; Fig. 6) and a conflict-gated rule that suppresses only where $c(b) < 0$.

What the surgery can and cannot do to first order. Write $g_{aa} = \nabla_{\theta_a} L_a$, $g_{va} = \nabla_{\theta_a} L_v$ for one block, so the surgery applies $d_a = g_{aa} + \alpha g_{va}$.

Proposition 1. *For a step of size η with the other tower fixed, $L_a(\theta_a - \eta d_a) = L_a(\theta_a) - \eta \|g_{aa}\|^2 (1 + \alpha\rho c) + O(\eta^2)$. Hence (i) the leaked gradient slows the audio objective only by $\rho(-c)_+$; (ii) the largest α preserving a fraction κ of native progress is 1 if $c \geq 0$ and $\min(1, (1-\kappa)/(\rho|c|))$ otherwise; (iii) $\min(1, \tau/\rho)$ is the worst case $c = -1$ of (ii).*

The proof is one line (Appendix E). The proposition is a diagnostic, not a guarantee: it is one step of one tower on a reconstruction loss and ignores the optimizer. What it gives is a scale. Where $\rho|c|$ is two orders of magnitude below one, no schedule can change the first-order audio dynamics, so any difference between schedules has to come from elsewhere and will be small.

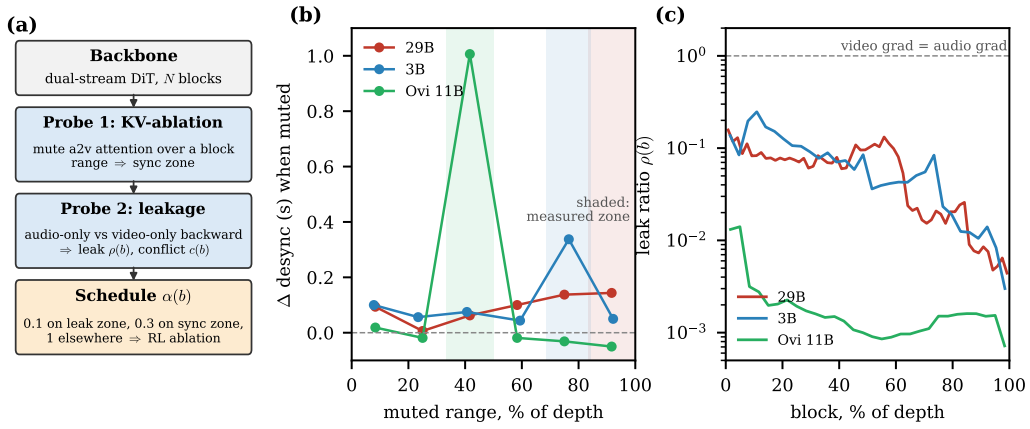

Figure 1: (a) The protocol: two probes and the schedule they yield. (b) Probe 1 on three backbones: change in desynchronization when the a2v direction is muted over a block range, by normalised depth; shaded band = selected zone. (c) Probe 2: audio-tower leak ratio $\rho(b)$ by normalised depth. The zones sit at different depths; Ovi's leakage is an order of magnitude below the in-house models'.

4 Experimental Setup

Backbones. (i) A 29B twin-tower AV transformer, 60 paired blocks, video width 4096 / audio 896, packed varlen flash attention, rotary cross-attention (Anonymous, 2026); LoRA r=16 on all attention and feed-forward projections of both towers (1680 modules); RL on 64 GPUs (60 training, 4 reward services), K=10, 12 prompts per epoch, 3000 bilingual prompts. (ii) A 3B sibling with 32 blocks and the same cross-attention; same recipe on 32 GPUs. (iii) Ovi (Low et al., 2025), 11B, twin 30-block towers with per-block bidirectional fusion, from an independent codebase; official checkpoint, LoRA r=16 on both towers, K=8, 832 prompts, 8 GPUs ($\approx$ 55 GPU-hours per 50-epoch arm). The reward stack (Appendix B) is HPSv3 and VideoAlign on video; AudioBox, CLAP and the ASR reward on audio; Synchformer, onset alignment and a face-gated lip-sync score broadcast to both. On Ovi the stack is reduced to Synchformer, ASR and CLAP.

Surgery arms. On the 29B backbone: No surgery ($\alpha \equiv 1$); Legacy, the schedule our production system had actually been running (0.1 on 1–5, 0.3 on 26–31: a window correctly rescaled for a 32-block predecessor and then carried over unchanged when the family grew to 60 blocks); Index-copy (published indices, 40–47); and Measured (1–10, 50–59). Each ran 50 epochs on 32 GPUs from a production LoRA that had been trained for 218 epochs under Legacy; No surgery and Measured were then repeated from the SFT base with no warm start. On the 3B backbone: No surgery, a proportional rescale of the LTX-2 window (27–31) and Measured (22–26), 100 epochs from the SFT base. On Ovi: seven arms — no surgery, a transferred LTX-2 schedule (20–23, rescaled from an assumed 60-block source, hence not the proportional 25–29 of Table 1), Measured (block 1 at 0.1, 10–14 at 0.3), the closed form $\min(1, \tau/\rho)$, shallow-only, conflict-gated, and a second training seed of Measured. Every arm has one training seed unless stated.

Evaluation. WER is Whisper-large-v3 against the scripted text, three sampling seeds per prompt, 50 steps with CFG, on two public prompt sets: 100 Harvard sentences (IEEE Subcommittee on Subjective Measurements, 1969) in five fixed speaker templates, and the 64-prompt `human_language` subset of VABench (Hua et al., 2025) (VABench-64; 63 prompts carry a script), which does not overlap the 3000 RL training prompts. Per-prompt values are seed means; comparisons are paired Wilcoxon tests (one-sided for RL-vs-base, two-sided otherwise) with bootstrap 95% CIs, unadjusted; Holm-adjusted values are in Appendix F. Two ASR judges not used as rewards (wav2vec2-large-960h, whisper-medium) re-score the main pairs. Video and audio quality use the official JavisBench-mini suite (1000 prompts, 16 metrics).

Table 1: Synchronization zone per backbone (probe 1; 32 prompts; one-sided paired Wilcoxon, muting increases |desync|; worse/better = videos whose desync increased/decreased; Δ in seconds; N = blocks), where the zero-cost heuristics would place it, and probe-2 leakage ($\rho_{\max}$: audio-tower group). On the three measured backbones no other a2v range is significant at $p < 0.05$ except the adjacent 40–49 decade on the 29B model (p=0.031); LTX-2 is discussed in Appendix C. n/a: heuristic not applicable (block 47 does not exist, or the model is the source itself); n/r: not reported. [§]30 audio slots, 40 video blocks; probes only.

Backbone	N	Zone	Depth	Index copy	Prop. rescale	Δdesync, w/b	p	$\rho_{\max}$
LTX-2 (reported[1])	48	40–47	83–98%	n/a	n/a	n/r	n/r	> 1
LTX-2 (ours, probes)	48	24–39[†]	50–81%	40–47 ×	40–47 ×	+0.38, 19/4	0.003	0.1–0.26
29B (ours)	60	50–59	83–98%	40–47 ×	50–59 ✓	+0.14, 7/1	0.023	0.16
3B (ours)	32	22–26	69–81%	n/a	27–31 ×	+0.34, 17/4	0.0035	0.25
3B, retrained	32	22–26	69–81%	n/a	27–31 ×	+0.58, 20/6	0.0011	0.28
Ovi (public)	30	10–14	33–48%	n/a	25–29 ×	+1.01, 26/1	5×10^{-6}	0.014
MOVA (public, probes)	30[§]	15–19	50–66%	n/a	25–29 ×	+0.31, 15/4	0.003	0.07–0.2

5 Results I: What Transfers Across Backbones

Zones sit at different depths. Fig. 1b and Table 1 give the probe-1 result. On each backbone exactly one a2v range degrades synchronization when muted, and it is in a different place: the last decade on the 29B model (coarsely; the adjacent decade is also significant, and the final one is chosen because it alone also degrades speech), blocks 22–26 on the 3B, and the middle of the network on Ovi, where muting blocks 10–14 desynchronises 26 of 32 videos and no other range reaches $p < 0.4$. A fifth architecture, the public MOVA (OpenMOSS, 2026), probed without RL, has its zone at 50–66% (Appendix A). On LTX-2 itself, the model whose schedule started all this, probe 1 does not find the published window: in four independent runs (three RL-regime seeds and one default-regime run) blocks 24–31 degrade synchronization every time and 32–39 in two, while 40–47 never does ($p \geq 0.05$; Appendix C). The reference reports its ablation in five coarse bins whose two deepest contain our zone, so its constant may have been read off a bin. Probe 2 on LTX-2 does reproduce leakage above one at the a2v key/value input, where the reference hook acts, but not on the audio-tower parameters ($\rho \leq 0.26$): the two definitions agree on where leakage enters and disagree on whether it dominates what the surgery protects. Without an RL arm on LTX-2 we do not count it as a fourth measured line. Copying the LTX-2 indices misses the zone on the 29B and is undefined on the 3B and Ovi, which end before block 40; proportional rescaling happens to land on the zone on the 29B (40–47/48 $\rightarrow$ 50–59/60) and misses on the 3B and Ovi.

Zones are invariant within a model line. Schedules measured at one SFT checkpoint were used unchanged for two 29B weight soups. A retrained 3B line — different pretraining seed, a different single-stage SFT recipe, a 12-domain soup and an added image-conditioning pathway — has the identical zone (Table 1, row 4). On Ovi, probe 1 on the RL-tuned checkpoint finds the zone unchanged (+1.16 s when muted, 29/32 videos). The zone is a property of the model line; one measurement per line is enough.

Leakage sits in the cross-attention groups and never dominates the audio tower. Probe 2 (Fig. 1c) records the ratio separately for the audio tower and for the a2v cross-attention parameters, and they behave differently. On the audio tower the leak ratio is 0.06–0.13 on the in-house models, decaying to 10^{-2}, and peaks at only 0.014 on Ovi: video supervision never overwhelms the native audio gradient there. On the a2v cross-attention groups, the parameters the scale actually touches, the ratio exceeds 1 on a minority of blocks (14 of 59 on the 29B, peaking at 53 at block

[1]Constants of Zhang et al. (2026); its ablation is a five-bin figure without per-video statistics and its ρ computation is not released. Our ρ is on LoRA gradients of the audio tower and not numerically comparable; at the a2v key/value input, where the reference hook acts, our LTX-2 ratio is 1.1–1.9 over blocks 0–10, so the published "> 1 in shallow blocks" holds at the input and is attenuated to ≤ 0.1 at the audio-tower parameters (Appendix C). [†]24–31 is significant in all four probe-1 runs, 32–39 in two, 40–47 in none; the quoted Δ and p are for 24–31 in the default sampling regime. No RL arm was trained on LTX-2 or MOVA.

Table 2: Surgery ablation, VABench-64 WER (64 prompts $\times$ 3 sampling seeds; one training seed per arm). 29B warm start: final epoch-50 and best-by-training-reward checkpoints; 29B clean start: from the SFT base; 3B: epoch 100 and best-by-reward checkpoint (epoch 15 on all three arms) from the SFT base; Ovi: best checkpoint of 50. p is two-sided paired vs. the respective base (final checkpoint where two are listed). Harvard WER: final / best where both exist. Numbers in parentheses are best epochs. Bold: significant degradation.

Backbone / start	Schedule	WER final	WER best	p vs base	Harvard final / best
29B, warm (base 0.193)	NO SURGERY	**0.228**	0.207	0.024	0.133 / 0.162
	LEGACY	0.180	0.183	0.19	0.161 / 0.168
	INDEX-COPY	0.195	0.206	0.73	0.121 / 0.111
	MEASURED	0.204	0.181	0.48	0.187 / 0.175
29B, clean (base 0.235)*	NO SURGERY	**0.326**	**0.317** (41)	0.038	0.234 / 0.224
	MEASURED	0.175	0.171 (43)	0.002	0.205 / —
3B, clean (base 0.179)	NO SURGERY	0.132	0.172 (15)	2×10^{-4}	0.223 / 0.249
	PROP. RESCALE	0.143	0.158 (15)	0.009	0.202 / 0.275
	MEASURED	0.135	0.157 (15)	0.001	0.269 / 0.210
Ovi, clean (base 0.186)	NO SURGERY	0.147	0.148	0.018	0.036
	TRANSFERRED	0.144	0.144	0.006	0.038
	MEASURED	0.139	0.145	0.003	0.035
	MEASURED, seed 2	0.138	0.148	0.007	0.039

*ep. 20/30/40/50: NO SURGERY .342/.275/.343/.326; MEASURED .212/.196/.183/.175.
Harvard bases: 29B warm .164, clean .251; 3B .325; Ovi .053.
Two sampling seeds instead of three: ep. 40, 29B clean Harvard best, 3B MEASURED best.

58, inside that model's measured zone) and on most blocks of the 3B (median 2.3–2.7). Pooling the groups hides this because the audio-tower gradients are one to two orders of magnitude larger, so we report them separately. The absolute norms are small in the deep blocks where the largest ratios occur, which is why the quantity that governs the audio step is $\rho|c|$ of Proposition 1 rather than ρ alone. On Ovi, where both towers fit on one device, the conflict cosine is small and *positive* over the first 22 audio blocks ($+0.005$ to $+0.035$; block 1: $+0.009$, 95% CI $[+0.003, +0.016]$), so rule (ii) leaves the audio tower unsuppressed; the only pronounced conflict on any backbone is in the mirror direction, at Ovi's penultimate video block ($c = -0.11$). The mirror ratio ρ' (audio loss into video parameters) exceeds 1 on all three backbones, up to 35 on the 29B, and every published schedule leaves that direction unscaled (Fig. 7, Appendix D).

6 RESULTS II: WHEN DOES SURGERY HELP

Table 2 collects every arm; Appendices G and I give per-seed values and all pairwise tests.

On the 29B backbone the surgery is necessary. From the warm start, NO SURGERY is the only arm that ends worse than it began (WER $0.193 \to 0.228$, p=0.024, CI $[+0.003, +0.068]$): it peaks at epoch 8 and decays while every surgery arm stays within ±0.013 of the base, and the one between-arm contrast that reaches significance is LEGACY vs. NO SURGERY (p=0.0014). The warm start is a confounder, since the base LoRA was trained under LEGACY and moving away from that schedule could itself cost, so we repeated the two arms from the SFT base. The effect is larger there. After 20 epochs NO SURGERY has raised WER from 0.235 to 0.342 (+46%, p=0.0008, CI $[+0.01, +0.20]$); 19 of 63 prompts regress by more than 0.2, the count at WER ≥ 0.5 rises from 9 to 15, and the training-side speech reward oscillates without trend. The arm never recovers: 0.275 at epoch 30, 0.343 at 40, 0.326 at 50 (+39%, p=0.038; per-seed 0.228/0.264/0.485, the last driven by babbling, the SFT base's own worst-seed failure). The checkpoint the training reward selects (epoch 41) has WER 0.317, 35% worse than the base: without surgery the total reward improves while speech degrades, which is the video-into-audio interference the surgery is meant to block, visible at the level of model selection. Two recognisers that were not used as rewards see the same degradation, larger if anything (+23/+40% final, +40/+59% best; Table 5).

The matched MEASURED arm from the same SFT base, same budget and same reward stack moves the other way: WER falls at every checkpoint (0.212/0.196/0.183/0.175 at epochs 20/30/40/50), ends 25% below the base (p=0.002, CI $[-0.098, -0.024]$, Holm $p \leq 0.012$; per-seed 0.158/0.173/0.194), improves 31 of 63 prompts against 9, and cuts the count at WER ≥ 0.5 from 9 to 5. Its best-by-reward checkpoint (epoch 43, 0.171) is also its best in speech, so reward and speech move together here. The gap between the two arms is -0.151 final-vs-final and -0.146 best-vs-best ($p < 0.001$), the largest contrast in the study and the one that isolates the schedule as the cause. On Harvard, MEASURED reaches 0.205 against the SFT base's 0.251 while NO SURGERY ends at 0.234 (p=0.78): on the easier public sentences RL without surgery neither helps nor hurts, so the damage is specific to the harder VABench speech. Fifty epochs is a quarter of the production budget; the production run from the same SFT base with MEASURED continues to 0.122 at epoch 139 (Sec. 7), whereas none of the four NO SURGERY checkpoints comes near the base. The contrast widens with budget.

On Ovi and on the 3B backbone no schedule is distinguishable from no surgery. On Ovi every arm improves on the base (WER 0.186 $\rightarrow$ 0.138–0.158, desync 0.41 $\rightarrow$ 0.29–0.37 s), the second training seed of MEASURED reproduces this, and none of the 21 pairwise between-arm tests on WER or desynchronization reaches $p < 0.05$ (smallest 0.088). On Harvard, where Ovi's base is already at 0.053, all four arms improve it by 27–34% ($p \leq 0.036$) and MEASURED vs. NO SURGERY gives p=0.68; under the two independent judges every arm improves on the base under wav2vec2 and all but MEASURED under whisper-medium, with no between-arm contrast below p=0.07 (Appendix I). This is what Proposition 1 predicts for $\rho \leq 0.014$ and $c > 0$; the closed-form and conflict-gated arms in particular match MEASURED within $p \geq 0.56$. A non-detection at 64 prompts and one training seed is not an equivalence proof, and we do not treat it as one. On the 3B all three arms improve on the base by 20–26% on VABench-64 and are pairwise indistinguishable ($p \geq 0.52$); on Harvard they all improve but MEASURED is now the worst (0.269 vs. 0.202–0.223, $p \leq 0.001$). No surgery is never worse than the probe-derived schedule on this backbone, and the two independent judges agree: NO SURGERY improves on the base under both ($p \leq 0.013$) and no between-arm contrast reaches $p < 0.1$ (Table 5). On all three 3B arms the reward-selected checkpoint (epoch 15) is worse in speech than epoch 100 (0.157–0.172 vs. 0.132–0.143), a second case of the total reward not tracking intelligibility.

Fine placement is not resolvable at this budget. Among the three 29B surgery schedules, WER differences are 0.01–0.02 and change sign between the final and the best checkpoint; on Harvard the ranking reverses — INDEX-COPY, middling on VABench-64, is best (0.111, p=0.002 vs. base) and MEASURED, best on VABench-64 at its best checkpoint, is worst (0.187). Training reward orders the arms as calibration would predict (NO SURGERY < LEGACY < INDEX-COPY < MEASURED), but with one training seed each that is a hypothesis. A schedule that is best on one prompt set and worst on another is not telling us about the schedule. Probe 1 locates cross-modal sensitivity reliably; it does not follow that suppressing the backward signal there gives a better policy, and our data do not show that it does.

What predicts necessity. Fig. 2 plots the effect of removing the surgery against the maximum leak ratio. The two extreme backbones behave as the ratio suggests: ρ=0.16 and a large, reproducible degradation on the 29B; ρ=0.014 and nothing on Ovi. The 3B backbone, with ρ=0.25 and no effect, shows that the ratio is not the whole story, and neither is the absolute leaked gradient: the 3B audio tower receives a *larger* video-loss gradient than the 29B one in absolute terms too (summed over blocks, 1.27 vs. 0.48; peak block 0.137 vs. 0.021; Appendix D). What differs is the receiving tower: $10\times$ fewer parameters, hence a stronger RL signal per parameter. The hypothesis consistent with our measurements is about recovery rather than injury: a small audio tower re-learns intelligible speech from the routed reward within the budget and a large one does not. It can be tested by training the 29B arm longer or the 3B arm with a weaker speech reward. The empirical rule we are willing to state is conservative: a leak ratio two orders of magnitude below one means the surgery is inert; a ratio near 0.2 on a large backbone has meant it is necessary; in between, ablate.

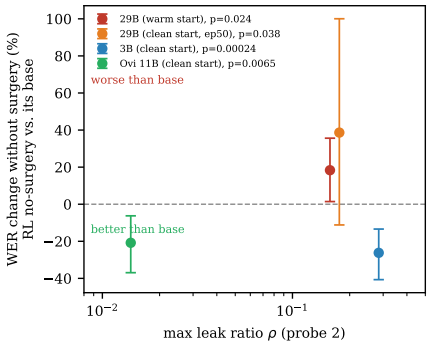

Figure 2: Relative WER change of the NO SURGERY arm vs. its base (VABench-64, 3 seeds, bootstrap 95% CI) against the maximum audio-tower leak ratio (probe 2): harmful on the 29B from both starts, harmless on Ovi, and no effect on the 3B despite strong leakage.

Table 3: WER of SFT bases and RL policies under the reward judge (Whisper-large-v3) and two independent judges, on the public Harvard-100 protocol and the VABench-64 speech subset (mean over prompts of three-seed means; one-sided paired Wilcoxon p vs. base).

		Whisper-L3 (reward)			wav2vec2			Whisper-M		
Backbone / protocol	RL ckpt	base	RL	p	base	RL	p	base	RL	p
29B Harvard-100	ep. 216	0.251	**0.180**	0.005	0.200	0.157	0.013	0.161	0.137	0.16
29B VABench-64	ep. 139	0.235	**0.122**	6×10^{-5}	0.272	0.207	2×10^{-4}	0.196	0.134	0.003
3B Harvard-100	ep. 100	0.325	**0.211**	6×10^{-7}	0.220	0.172	9×10^{-5}	0.209	0.160	0.001
3B VABench-64	ep. 100	0.179	**0.140**	0.005	0.246	0.224	0.026	0.163	0.150	0.21
Ovi VABench-64	best	0.186	**0.145**	0.003	0.357	0.342	0.028	0.186	0.161	0.23
Ovi Harvard-100	best	0.053	**0.035**	0.002	—	—	—	—	—	—

7 RESULTS III: SPEECH AND AUDIO–VIDEO QUALITY

The speech objective. Prompts for human-centric AV generation often contain a literal script (in our data, between markers), which makes intelligibility checkable: we transcribe the generated audio with Whisper-large-v3 and use $r = 1 - \min(\mathrm{WER}, 1)$ as a reward, routed to the audio branch only; prompts without a script get a neutral value that cancels in the group-normalised advantage, and scripted prompts are oversampled $1.5\times$. Routing to audio only is a choice, not a necessity (the video branch can affect acoustics through cross-attention); a symmetrically routed variant lowered video rewards in early runs without speeding up WER gains. We did not run an arm without the ASR term, so the numbers below measure the recipe, not the term on its own.

Speech. Under the reward judge, RL lowers WER by 28% on Harvard and 48% on VABench-64 for the 29B backbone, by 35% and 21% for the 3B at epoch 100 (52% and 29% at its best checkpoint, epoch 215; Appendix G), and by 22% for Ovi (34% on Harvard, where the base is already at 0.053); a second training seed reproduces the Ovi result. Under the two independent judges the direction holds in every pair and is significant in eight of twelve, but the size shrinks to 15–31% on the 29B and 4–14% on Ovi; we report both. On VABench-64 the base is strongly seed-sensitive (per-seed WER 0.19–0.32) and the RL policy is not (0.09–0.14; Fig. 10); the same happens on Ovi and, more weakly, on Harvard. The failures the objective removes are near-silent speech, babbling past the end of the script, and slurring (Fig. 8). Training-window rewards improve together (visual preference $7.8 \rightarrow 9.4$, video–text alignment $1.36 \rightarrow 1.77$, synchronization flat; Fig. 9), so the routed speech reward does not cost video quality.

External AV quality. JavisBench-mini shows the same pattern on all three backbones (Appendix H): FAD, audio, visual and motion quality, text–audio alignment and CLAP improve (29B: FAD $6.20 \rightarrow 5.75$, visual quality $2.74 \rightarrow 2.95$, motion $0.151 \rightarrow 0.266$); text–video alignment is

flat; FVD gets slightly worse ($+1$–6%), and the benchmark's desynchronization judge, which shares weights with our training-time reward, worsens by about 0.03 s on the 29B. Among the 29B ablation arms, MEASURED has the best FVD, visual and audio quality and desync at its final checkpoint. That is the one external signal for the probe-derived zones; we do not weight it above the speech results.

8 DISCUSSION AND LIMITATIONS

What we claim and what we do not. The probes locate cross-modal sensitivity and leakage reliably, and the zones follow no fixed heuristic across architectures. Whether the surgery matters depends on the backbone: necessary on the 29B from both starts, not detectably useful on the 3B and Ovi. We do *not* show that the measured zone gives a better policy than a transferred one; the Harvard reversal says the question cannot be answered at this budget. Not detecting a difference is not the same as showing equivalence; that would need a pre-registered margin and training-seed replication.

One training seed per arm. Every between-schedule comparison rests on one training seed; sampling seeds measure generation variance only. This also qualifies the surgery-versus-none result on the 29B, even though it is reproduced from two starting points with the same sign. The clean-start MEASURED arm improves on the base by 25% where the warm-start one stayed within ± 0.013; we read this as headroom (the warm base had already absorbed 218 RL epochs), but it is one run. All p-values are unadjusted; under Holm adjustment every RL-versus-base result survives, as does the clean-start MEASURED-versus-NO SURGERY contrast on the 29B ($p < 0.001$), and no other between-schedule comparison does (Appendix F).

Scope of the probes and the analysis. Probe 2 uses reconstruction losses, so ρ and c describe the architecture rather than a reward stack; the conflict cosine could only be measured on Ovi, and the 29B–3B difference is where it is most needed. Proposition 1 is local and ignores the optimizer and the KV-hook's approximation of g_{va} (Appendix E). Speech evaluation is single-speaker and short-form, and the speech term is not isolated from the rest of the recipe. The in-house weights are not public at submission; a release is scheduled before the review period, and Ovi, the probes and the Harvard protocol can be reproduced now.

9 CONCLUSION

The constants of the current AV-RL recipe are measurements of one backbone. Two probes recover them on any two-stream model in hours and show that the zones move with the architecture but not with the checkpoint. Controlled ablations on three backbones show that what matters is whether the surgery is applied, not where: on one backbone removing it leaves speech worse after 50 epochs and decouples training reward from intelligibility, while the calibrated schedule from the same start cuts the error by a quarter; on two others it changes nothing we can detect; nowhere can the fine placement be resolved. Leakage magnitude separates the extremes but does not settle the middle. As RL recipes pick up architecture-specific machinery, the measurement behind each constant, and a test of whether the constant matters, is worth more than the constant itself.

AI USE STATEMENT

Large language models were used as coding and writing assistants: for drafting and refactoring experiment orchestration code, literature search, and editing the prose. All experimental designs, measurements and claims were specified, verified and validated by the authors.

REPRODUCIBILITY STATEMENT

Probe implementations, the Harvard-100 prompts with speaker templates and WER scoring code, the statistics script behind every paired test, and the Ovi training and evaluation configurations will be released. Ovi experiments are reproducible end-to-end from the official checkpoint; the 29B and 3B backbones are scheduled for public release before the review period. Hyperparameters and routing: Appendix B; evaluation protocol: Sec. 4, Appendix F.

REFERENCES

Anonymous. Technical report of a 29b dual-stream audio-video diffusion model, 2026. Concurrent technical report; citation to be completed upon public release.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2016.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *ICLR*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. *arXiv preprint arXiv:2512.09299*, 2025.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP*, pp. 5325–5329, 2024.

IEEE Subcommittee on Subjective Measurements. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025c.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026a.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026b.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

OpenMOSS. MOVA: Open joint audio–video generation with dual wan towers. `https://github.com/OpenMOSS/MOVA`, 2026. Commit 1591b25, checkpoint MOVA-720p.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *arXiv preprint arXiv:2407.13734*, 2024.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Guohui Zhang, Xiaoxiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. Omninft: Modality wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.