Adding Voice Cloning to Text-to-Audio-Video Models with a Single Zero-Initialised Layer

Ivan Mikheev, Viacheslav Vasilev, Anna Dmitrienko, Alexey Letunovskiy, Ivan Kirillov, Kirill Chernyshev, and Denis Dimitrov

Kandinsky Lab, Moscow, Russia
{ivan.mikheev,viacheslav.vasilev}@kandinskylab.ai

Abstract. Text-to-audio-video models generate video and audio from text but lack speaker identity control. We add voice cloning via a single zero-initialized linear layer atop the audio backbone with short fine-tuning, conditioning on a reference clip (prepended latents + global speaker embedding). On 674 speaker–text pairs (30 speakers), our 5B model outperforms five baselines in speaker-identity similarity across three verification networks. Running only the audio path yields ∼30× speed-up while preserving voice-cloning behaviour.

Keywords: Text-to-Audio-Video generation · Voice cloning

1 Introduction

Text-to-audio-video (T2AV) diffusion models [7,12,13] generate video and audio from text but lack speaker control; text-to-speech (TTS) voice cloners [3, 17] copy timbre from a reference but are speech-only, and adding this to a pretrained T2AV model requires costly retraining or careful architectural surgery.

We present a minimal recipe for adding reference-conditioned voice cloning to pretrained T2AV models: two modifications – prepending reference latents to the audio stream and modulating audio with a global speaker embedding via FiLM [14] – fit naturally into the asymmetric audio–video DiT architecture and require only a single zero-initialized linear layer, so the augmented model starts as a functional copy of the backbone. We apply this to two models, K5 (5 B) and K5-LITE (0.6 B), with a short single-stage fine-tuning that preserves original generation quality. We introduce a three-encoder evaluation protocol on a 674-sample benchmark of unique-text speaker pairs, where our model significantly outperforms strong TTS baselines in speaker fidelity. We also demonstrate a practical inference variant that runs the audio path in isolation, yielding a ∼30× speed-up while retaining full voice-cloning behaviour.

2 Method

2.1 Base model

Our backbone is an asymmetric audio–video diffusion transformer (AV-DiT) with separate video and audio streams ($d_v = 1792$, $d_a = 896$), connected via

cross-modal attention inside each fused block. The architecture is based on the open-source CrossDiT from Kandinsky 5.0 [1]; the text condition is shared between streams, while audio latents are produced by an off-the-shelf neural audio VAE [5] at 44.1 kHz and video latents by a HunyuanVideo VAE [11]. Further architectural details are given in Appendix A.1. We experiment with two checkpoints trained on a large corpus of audio–video scenes: **K5** (5 B parameters) and **K5-LITE** (3 B total, whose audio sub-network accounts for 0.6 B). After fine-tuning, we denote them as **K6A_5B** (full audio–video inference) and **K6AV_LITE** (audio-only inference, evaluating only the 0.6 B audio sub-network).

2.2 Reference-conditioned voice cloning

We inject two complementary reference signals into the AV-backbone:

Prepended reference latents. For target latents $x_a = \mathrm{VAE}(a_{\mathrm{target}}) \in \mathbb{R}^{T_a \times d_a}$ and a 1.2–4 s reference a_{ref}, we encode the reference with the same VAE and prepend:

$$\tilde{x}_a = [\,\mathrm{VAE}(a_{\mathrm{ref}}) \,\|\, x_a\,] \in \mathbb{R}^{(T_{\mathrm{ref}}+T_a)\times d_z}.$$

Across the 32 fused decoder blocks, target tokens attend to reference tokens via the existing self-attention (no new layers). At inference, the diffusion update is restricted to the target portion of the composite sequence; the reference portion is held fixed so that the model is forced to keep its timbre consistent across all denoising steps.

Speaker FiLM. A global speaker embedding $e \in \mathbb{R}^{1024}$ from a frozen Qwen3-TTS encoder [9] is projected by a single new linear layer $W_{\mathrm{film}} : \mathbb{R}^{1024} \to \mathbb{R}^{2d_a}$ to scale and shift coefficients (γ, β), which modulate each target token via FiLM [14]:

$$h_a \leftarrow h_a \odot (1+\gamma) + \beta, \quad (\gamma, \beta) = W_{\mathrm{film}}(e).$$

Reference tokens are not modulated. W_{film} is zero-initialized, so $(\gamma, \beta) = (0, 0)$ initially and the augmented model output matches the base model on the first forward pass.

2.3 Classifier-free guidance with two directions

During training, we independently drop the text ir the reference signal (latents + speaker embedding) with probability 0.1 each, providing the four configurations for three-way split CFG at inference [8]:

$$\hat{\epsilon} \;=\; \epsilon_{\varnothing} + w_t\,(\epsilon_{\mathrm{text}} - \epsilon_{\varnothing}) + w_r\,(\epsilon_{\mathrm{full}} - \epsilon_{\mathrm{text}}), \tag{1}$$

where $\epsilon_{\varnothing}$, ϵ_{text}, and ϵ_{full} are unconditional, text-only, and full-conditioned predictions. The first guidance term controls prompt adherence, the second controls reference voice imposition strength, enabling separate trade-off of text vs. speaker fidelity at inference without retraining.

2.4 Training schedule

We fine-tune the entire model (base weights + W_{film}) in a single voice-aware stage. We use AdamW with 1×10^{-5} for base parameters and 5×10^{-5} for W_{film}; the reference window is sampled from speech-rich segments after the target. In 10% of steps, we feed unnoised audio/video latents to preserve the original AV distribution. That is, the model remains a T2AV generator first and a voice-cloner second. Implementation details are in Appendix A.

3 Experiments

3.1 Benchmark and metrics

We evaluate on a voice-cloning benchmark built from VCTK [18]: 674 samples from 30 speakers with unique texts (5-16 words). Detailed setup (pre-processing, duration control, and metric computation) is given in Appendix B.

We compare against five external baselines: Qwen3-TTS (1.7B and 0.6B) [9], XTTS-v2 [2], IndexTTS2 [19], and NAVA [10].

We report speaker-encoder cosine similarity (SECS) using three verification networks (ECAPA-TDNN [6], WavLM-SV [4], and Resemblyzer/GE2E [16]), computed both against the reference clip and against a centroid of four clips. Intelligibility is measured by WER%, CER% of Whisper transcriptions [15], and by $\mathbf{WER_0\%}$, the fraction of samples transcribed with zero word errors.

3.2 Main results

Table 1 reports the comparison on the full 674-sample benchmark. **K6A_5B** attains the highest speaker similarity on *every* one of the six SECS columns, outperforming all six competing systems on both the reference-based and the centroid-based scores across all three verification networks. The lead holds on the two most reliable encoders, WavLM-SV (0.944 vs. reference) and Resemblyzer (0.866), as well as on ECAPA-TDNN (0.766); averaged over the three networks the margin over the strongest external baseline (Qwen3-TTS 0.6B) is 0.041 and is confirmed by a paired Wilcoxon signed-rank test ($p < 10^{-89}$ against each competitor).

Qwen3-TTS and IndexTTS2 achieve low WER but lower SECS (clean, studio-like voice). Our models trade higher WER for stronger speaker fidelity, faithfully reproducing reference acoustics. This reflects the diffusion prior: it reconstructs fine reference details, occasionally hallucinating words on short/difficult prompts.

3.3 No regression of the base audio model

Adding voice-cloning conditioning does not degrade the base model's reference-free generation: run without a reference on 400 prompts, the fine-tuned model in fact improves on every objective axis (FAD, WER, CER, WER_0, CLAP) and matches perceptual naturalness (UTMOS), owing to the zero-init W_{film} that keeps the FiLM path near identity when no reference is present. A human side-by-side study confirms the same pattern; full numbers are in Appendix E.

Table 1: English VCTK benchmark (674 samples per model, 30 speakers, unique texts). The SECS columns are speaker-encoder cosine similarities (higher is better) for three verification networks (E: ECAPA-TDNN, W: WavLM-SV, R: Resemblyzer), reported both against the single reference clip (*vs. reference*) and against the centroid of the reference and three enrollment clips (*vs. centroid*). WER%/CER% are word/character error rates (lower is better) and $\text{WER}_0\%$ is the fraction of samples transcribed with zero word errors (higher is better). Best value per column in **bold**, second best underlined.

	SECS vs. reference			SECS vs. centroid					
Model	E	W	R	E	W	R	WER%	CER%	$\text{WER}_0\%$
K6A_5B (ours)	**0.766**	**0.944**	**0.866**	**0.770**	**0.951**	**0.878**	5.76	5.21	86.6
Qwen3-TTS 0.6B	0.678	0.936	0.840	0.711	0.946	0.864	0.81	0.36	**95.7**
Qwen3-TTS 1.7B	0.674	0.938	0.839	0.709	0.951	0.864	**0.74**	**0.17**	95.3
IndexTTS2	0.695	0.885	0.854	0.693	0.878	0.867	0.96	0.22	93.2
XTTS-v2	0.575	0.923	0.816	0.611	0.940	0.841	1.14	0.32	93.2
K6AV_LITE (ours)	0.640	0.847	0.826	0.630	0.846	0.843	5.08	3.38	69.4
NAVA	0.624	0.852	0.759	0.625	0.850	0.780	5.82	3.21	69.1

3.4 Fast Audio-only Inference from an AV Checkpoint

The asymmetry of the AV-DiT lets us run the audio path *in isolation* at inference without changing any weights, by short-circuiting the video sub-block of each fused decoder block and skipping the video VAE decoder (details in Appendix C). This drops the cost of one diffusion step from a 3.18 B-parameter joint pass to an effective $\sim$0.58 B audio-only pass, a $\sim$30$\times$ speed-up. The speed-up is not bought at the cost of voice cloning: in Table 1, K6AV_LITE preserves the reference timbre well, at a modest SECS drop of $\approx$ 0.09 vs. the full K6A_5B, which we read as the video signal acting as a mild regularizer on the audio path. In practice, K6AV_LITE enables quick auditioning of a reference across many prompts before committing to full K6A_5B rendering.

4 Conclusion

The proposed prepend+FiLM recipe is a simple and effective way to add voice cloning to pretrained T2AV models. Two immediate extensions are natural: the recipe should transfer to any asymmetric AV-DiT paired with a frozen speaker encoder, and the observed WER gap to dedicated TTS systems suggests adding a small text-fidelity loss without disturbing the speaker conditioning.

References

1. Arkhipkin, V., Korviakov, V., Gerasimenko, N., Parkhomenko, D., Vasilev, V., Letunovskiy, A., Vaulin, N., Kovaleva, M., Kirillov, I., Novitskiy, L., Koposov, D., Kiselev, N., Varlamov, A., Mikhailov, D., Polovnikov, V., Shutkin, A., Agafonova, J., Vasiliev, I., Kargapoltseva, A., Dmitrienko, A., Maltseva, A., Averchenkova, A., Kim, O., Nikulina, T., Dimitrov, D.: Kandinsky 5.0: A family of foundation models for image and video generation (2026), https://arxiv.org/abs/2511.14993
2. Casanova, E., Davis, K., Gölge, E., et al.: XTTS: A massively multilingual zero-shot text-to-speech model. In: Interspeech (2024)
3. Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A.: YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2709–2720. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/casanova22a.html
4. Chen, S., Wang, C., Chen, Z., et al.: WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. (2022)
5. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis (2025), https://arxiv.org/abs/2412.15322
6. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Interspeech (2020)
7. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), https://arxiv.org/abs/2601.03233
8. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
9. Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., Zhang, X., Zhang, P., Yang, B., Xu, J., Zhou, J., Lin, J.: Qwen3-tts technical report (2026), https://arxiv.org/abs/2601.15621
10. Ji, L., Wang, G., Wei, X., Yang, C., Liu, X., Zhang, Z., Wang, S., Sun, Y., He, J.: Native audio-visual alignment for generation (2026), https://arxiv.org/abs/2605.30073
11. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025), https://arxiv.org/abs/2412.03603
12. Li, Y., Si, H., Landi, F., Gallegos, P.O., Koutsoumpas, I., Vazquez, O.R.C., Fu, R., Guo, Q., Jin, X., Liu, S., Song, M.: 3mdit: Unified tri-modal diffusion transformer for text-driven synchronized audio-video generation (2025), https://arxiv.org/abs/2511.21780

13. Liu, H., Lan, G.L., Mei, X., Ni, Z., Kumar, A., Nagaraja, V., Wang, W., Plumbley, M.D., Shi, Y., Chandra, V.: Syncflow: Toward temporally aligned joint audio-video generation from text (2024), https://arxiv.org/abs/2412.15220
14. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: visual reasoning with a general conditioning layer. In: AAAI. AAAI'18/IAAI'18/EAAI'18, AAAI Press (2018)
15. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: ICML (2023)
16. Wan, L., Wang, Q., Papir, A., Moreno, I.L.: Generalized end-to-end loss for speaker verification. In: ICASSP (2018)
17. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: Neural codec language models are zero-shot text to speech synthesizers (2023), https://arxiv.org/abs/2301.02111
18. Yamagishi, J., Veaux, C., MacDonald, K.: CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92) (2019). https://doi.org/10.7488/ds/2645, https://doi.org/10.7488/ds/2645
19. Zhou, S., Zhou, Y., He, Y., Zhou, X., Wang, J., Deng, W., Shu, J.: Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech (2025), https://arxiv.org/abs/2506.21619

A Implementation details

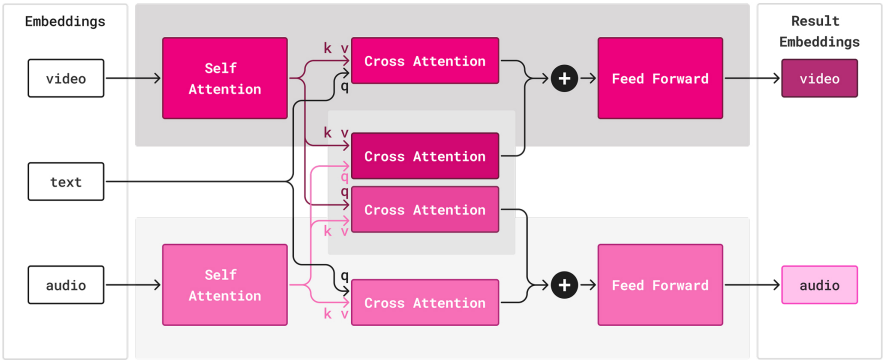

Fig. 1: Decoder transformer block for our AV-DiT architecture. Audio and video streams are fused using cross-attention. Each stream is based on the CrossDiT architecture [1].

A.1 Base architectures

K5 and K5-LITE share the same asymmetric AV-DiT backbone described in Section 2.1 (Figure 1). Both consist of 32 fused decoder blocks, with a video stream of width $d_v = 1792$ (heads of dimension 64) and an audio stream of width $d_a = 896$ (heads of dimension 64). A 2-block text encoder consumes the prompt and feeds it to both streams. K5 uses the full 32 fused blocks, K5-LITE a reduced configuration with the same block topology but smaller text and output layers, totaling $\approx 3\,\mathrm{B}$ parameters, of which the audio stream (the part evaluated in the audio-only inference mode) accounts for only $\approx 0.6\,\mathrm{B}$.

A.2 Reference window sampling

At training time, given a training clip, we sample the target audio from the first slice of the parquet record (typically 5 seconds of audio aligned with the video latents). The reference window is drawn from the *remaining* audio of the same clip, with a buffer of 0.5 s between target and reference to avoid trivial copy. Speech references are sampled with duration in $[1.2, 2.5]$ s from a 12 s search window, non-speech references with duration in $[1.5, 3.5]$ s from a 3.5 s search window. A minimum-RMS and a maximum-clipping-ratio quality filter discards silent or saturated references.

A.3 Hyperparameters

- Optimizer: AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-3}, max grad norm 1.0.
- Base LR: 10^{-5}; LR multiplier for the new linear layer W_{film} and its biases: $5\times$.
- Warmup: 8000 steps, constant afterwards.
- Reference-conditioning drop probability: 0.1; text-conditioning drop probability: 0.1.
- Modality-clean schedule: 10% of steps run with the video latent clamped to zero (audio-only loss) and 10% with the audio latent clamped to zero (video-only loss), to retain joint AV behavior throughout fine-tuning.
- Default inference: 50 diffusion steps, split CFG with $w_t = 5$ and $w_r = 4$, fixed seed.

A.4 Speaker encoder

The frozen speaker encoder used to compute the FiLM [14] embedding is a publicly released Qwen3-TTS speaker network [9] operating on 24 kHz audio. We run it once per training sample on CPU in a background worker and only transmit the resulting 1024-dim vector to the training step.

B Benchmark details

The benchmark uses VCTK [18] at native 48 kHz studio quality. Each of the 30 speakers is assigned its own sentences (100% unique texts, 5–16 words) to avoid text overlap. Each sample provides one reference clip, three additional enrollment clips of the same speaker, and a held-out target text.

For fair comparison, we feed each reference to every model in its own native sample rate, apply identical pre-processing: silence trimming, peak-normalisation to -3 dBFS, downmixing to 16 kHz mono, and evaluate all systems at the same 16 kHz rate. Our diffusion models have no explicit target-length conditioning; we set the generated duration adaptively: seconds = words/2.6 + 1.3, clamped to $[4, 12]$ s. The TTS baselines determine their own length.

The three verification networks are ECAPA-TDNN [6], WavLM-base-plus-sv [4], and Resemblyzer (GE2E) [16]. For each network we compute two similarities that differ only in what the generated utterance is compared against. The *vs. reference* score is the cosine similarity between the generated embedding and the *single* reference clip given to the model. The *vs. centroid* score instead compares the generated embedding against the *centroid* of four clips of the same speaker (the reference plus three enrollment clips), which reduces the variance of a single-utterance reference and gives a more stable estimate of the target speaker identity. Intelligibility is measured by word and character error rates **WER%**/**CER%** of Whisper transcriptions [15], and by $\mathbf{WER}_0\mathbf{\%}$, the fraction of samples transcribed with zero word errors.

C Audio-only inference details

Because the AV-DiT is asymmetric, the same fine-tuned checkpoint can be evaluated with the video stream disabled, without any weight changes (Sec. 3.4). Concretely, inside each of the 32 fused decoder blocks the video sub-block is short-circuited: its 1792-dim video self-attention and FiLM modulation are never computed, while the audio sub-block, the speaker FiLM layer, and the audio classifier-free-guidance loop keep operating exactly as in training. The video VAE decoder is likewise skipped. This reduces the per-step forward from a 3.18 B-parameter joint-modality pass to an effective $\sim$0.58 B audio-only pass, giving the $\sim$30$\times$ speed-up reported in the main text. The audio-only variant K6AV_LITE trades a small amount of speaker similarity (≈ 0.09 in SECS averaged over the three encoders) for this speed-up; we attribute the gap to the loss of the video stream, which otherwise acts as a mild regularizer on the audio path.

D Additional analysis

We carried out a more extensive factor study to identify which inference-time and data-time choices matter most for voice cloning quality. The following findings are based on a 82-run sweep across a held-out subset of the benchmark.

Reference length. Speaker similarity grows monotonically with reference length from $\sim$ 1 s up to a plateau at 4 s, beyond which returns are negligible. Below 2 s the Resemblyzer SECS drops sharply, suggesting that the speaker encoder rather than the diffusion model is the bottleneck on very short references.

Reference-guidance weight. The optimum of the second CFG weight w_r (Eq. 1) is in the range 4–6. Below 3 the reference is under-imposed; above 7 we observe a small over-smoothing of the audio that reduces UTMOS.

Number of diffusion steps. The mean SECS is essentially saturated by 30 steps; moving from 30 to 60 adds < 0.02 in mean SECS and is not worth the doubled wall-clock cost for routine use.

Language matching. Generating from a prompt whose language differs from the reference language is possible but lowers the mean SECS by about 0.05 on average. We therefore recommend matching prompt and reference languages whenever possible.

Reference pre-processing. A simple de-noise / VAD pre-processing of the reference recording is helpful on noisy or amateur captures but is harmful on already-clean studio references, where it can slightly hurt SECS. A safe default is to skip pre-processing when the reference already has UTMOS $\gtrsim 3$.

Table 2: No-regression check: *relative* change of the base T2AV model *after* reference-aware fine-tuning, evaluated in the reference-free regime on 400 prompts (30 speakers, identical seeds/texts). A positive change is an improvement for ↑-metrics and a negative change is an improvement for ↓-metrics; every metric moves in the improving direction.

Metric	Relative change
FAD vs. real speech ↓	−30.6%
WER ↓	−46.0%
CER ↓	−47.0%
WER_0 (perfect samples) ↑	+113%
CLAP text–audio ↑	+29.4%
UTMOS naturalness ↑	+0.6%

E No-regression details

We verify that adding voice-cloning conditioning does not degrade the reference-free generation ability of the pretrained audio backbone. We compare the base T2AV checkpoint against *itself after our reference-aware fine-tuning*, running *both* models in the reference-free regime (text→audio only, no reference clip) on 400 prompts covering 30 speakers, with identical seeds and texts. A regression would show up as the fine-tuned model moving *away* from real speech or losing intelligibility. We measure Fréchet Audio Distance (FAD) against real speech, word/character error rates (WER/CER), the fraction of perfectly transcribed samples (WER_0), CLAP text–audio alignment, and UTMOS naturalness.

Table 2 reports the *relative* change of the fine-tuned model with respect to the base model. Rather than regressing, the fine-tuned checkpoint improves on every objective axis: it is substantially closer to the distribution of real speech (FAD), roughly halves the transcription error and more than doubles the fraction of perfectly transcribed samples, aligns better with the prompt (CLAP), and matches perceptual naturalness (UTMOS). We attribute this to the zero-init design of W_{film} (Sec. 2.2): when the reference is absent the FiLM path stays close to identity, so the augmented model remains a strict functional superset of the base T2AV generator while benefiting from extra in-domain updates.

E.1 Human side-by-side study

In addition to the objective no-regression check (Sec. 3.3, Table 2), we run a human side-by-side (SBS) study to confirm that reference-aware fine-tuning does not degrade the perceived quality of reference-free generation. We compare the base audio model against *itself after our reference-aware fine-tuning* on 100 held-out text-to-audio prompts, with both models generating *without* a reference recording, so that any regression would show up as a lower win rate for the fine-tuned model. Annotators rate each pair on prompt following, technical quality, and speech/aesthetic quality as either a win for one of the two models or a draw. Table 3 shows that the fine-tuned checkpoint scores marginally higher win rates

Table 3: SBS human evaluation on 100 held-out text-to-audio prompts, base audio model vs. the same model after reference-aware fine-tuning (both generate *without* a reference). Numbers are percentages of pair-wise judgements per axis.

	Prompt follow.	Technical quality	Speech/Aesth. quality
Base wins	9	8	11
Fine-tuned wins	14	23	17
Both good	78	75	78
Both bad	9	4	4
Base win rate	48	43	47
Fine-tuned win rate	**52**	**57**	**53**

(52–57%) on all three axes, consistent with the objective results. A smaller SBS study on the full T2AV pipeline (reference-free) shows the same pattern, so the "strict superset" property carries over from audio-only to full multimodal.