INSTRUCTTTS-OPD: ON-POLICY DISTILLATION FOR MULTI-OBJECTIVE INSTRUCTION-FOLLOWING TTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-following text-to-speech (Instruct-TTS) renders a target transcript in the manner described by a free-form natural-language instruction, and a good rendition must succeed on three fronts at once: following the instruction, pronouncing the transcript correctly, and sounding natural. Reinforcement learning has become the standard way to post-train such generative models toward human preferences, and TTS systems increasingly adopt it. When the three objectives are optimized together under a single scalar reward, typically with group relative policy optimization (GRPO), training tends to be unstable: the objectives compete, and the shared update drifts toward whichever one is easiest to improve, producing a seesaw among them. We instead adapt on-policy distillation (OPD), which fuses several single-objective teachers by matching their velocity fields on the student's own rollouts. Realizing this for Instruct-TTS first calls for a sound reward on each axis, which we analyze carefully: a distilled Qwen-Omni proxy for the costly Gemini judge, dialect-free data for the WER reward, and DNSMOS in place of the language-biased UTMOS. We then train three single-objective teachers with GRPO and fuse them with OPD, routing each rollout to the instruction-following or WER teacher by its data domain and letting every rollout additionally follow the MOS teacher for quality. This lets instruction following and intelligibility improve together without a trade-off, but the quality target conflicts with instruction following, so following it additively reintroduces the seesaw. We therefore propose conflict-aware velocity fusion, which projects the quality target off the direction that opposes the primary objective before fusing, and show that it markedly reduces the instruction-following/quality trade-off. The fused model improves over scalar-reward GRPO on all three objectives at once. Code is available at `https://anonymous.4open.science/r/InstructTTS-OPD`.

1 INTRODUCTION

Text-to-speech (TTS) has advanced to the point where the synthesized voice is often hard to tell from a real recording (Anastassiou et al., 2024; Du et al., 2025; Hu et al., 2026a; Xu et al., 2026b; Liu et al., 2026c), and attention is shifting from merely reading a transcript to controlling *how* it is read (Li et al., 2026a; Yang et al., 2024). Instruction-following TTS (Instruct-TTS) captures this goal: given a target transcript together with a free-form natural-language instruction, such as "*read this in a trembling, sobbing voice*" or "*speak the following line in Cantonese, sounding impatient*," the model must render the transcript in the manner described (Guo et al., 2023; Yang et al., 2024). A good rendition has to succeed on three fronts at once: it must follow the instruction (instruction following, IF), pronounce the transcript correctly (intelligibility, measured by word or character error rate, WER/CER), and sound natural (quality, measured by mean opinion score, MOS). Among the three, instruction following is primary, as it distinguishes Instruct-TTS from ordinary TTS.

TTS models are increasingly post-trained with reinforcement learning (RL) to align their outputs with human preference (Ma et al., 2026b; Zhang et al., 2024; Wang et al., 2026b; Yang et al., 2026), and the dominant recipe is group relative policy optimization (GRPO) (Shao et al., 2024; Liu et al., 2026b) with a scalar reward. When several objectives must be optimized together, they are collapsed into a single number, and this scalarization discards directional information: when two objectives disagree on how the output should change, their contributions partially cancel, and a single scalar cannot express which trade-off to make. Optimization then drifts toward whichever objective is

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.

easiest to improve, so a softer objective comes to dominate the shared update while the competing ones regress. The result is the familiar seesaw among the three objectives, and reweighting the scalar only shifts the balance point rather than removing the tension.

A promising alternative comes from image generation, where on-policy distillation (OPD) fuses several single-objective expert teachers by matching their velocity fields on the student's own rollouts (Fang et al., 2026), giving a dense, per-step supervision signal in place of a single scalar. Bringing this idea to Instruct-TTS is not immediate, because the task carries characteristics that the image setting does not, and it first calls for a careful choice of reward along each of the three axes. For instruction following, the reliable judge is Gemini (Comanici et al., 2025), but querying it during training is prohibitively expensive; we therefore distill a Qwen-Omni (Xu et al., 2025) reward model that closely tracks the Gemini judgment and serves as a low-cost replacement. For intelligibility, we find that the WER teacher must be trained on dialect-free data: otherwise the model learns to lower WER by not speaking the requested dialect, which in turn drags down IF. For quality, we show that UTMOS (Saeki et al., 2022) is a poor choice because it carries a strong bias between Chinese and English, which pushes the model toward a foreign-accented reading of Chinese, and we adopt DNSMOS (Reddy et al., 2022) instead.

With these rewards in place, we train three single-objective teachers with GRPO and fuse them through OPD. We route each rollout to the IF or WER teacher by its data domain, and let every rollout additionally follow the MOS teacher for quality. Routing the two primary objectives to disjoint utterances lets instruction following and intelligibility improve together without a trade-off. Quality, however, is applied to every rollout and, as we show, conflicts with instruction following, so following it additively reintroduces the seesaw. We therefore propose conflict-aware velocity fusion, which projects the quality target off the direction that opposes the primary objective before fusing, a per-frame geometric operation on the distillation signal that a scalar reward cannot express. This markedly reduces the instruction-following/quality trade-off, and the fused model improves over scalar-reward GRPO on all three axes at once.

The main contributions of this work are summarized as follows:

- **Reward analysis for Instruct-TTS.** We identify a sound reward along each axis: a distilled Qwen-Omni proxy for the costly Gemini judge, a dialect-free WER reward, and DNSMOS in place of the language-biased UTMOS.
- **On-policy distillation for multi-objective Instruct-TTS.** We fuse three single-objective GRPO teachers with OPD, routing each rollout to one teacher by its data domain, and surpass scalar-reward GRPO on all three axes.
- **Conflict-aware velocity fusion.** Because the quality target conflicts with instruction following, we project it off the opposing direction before fusing, a per-frame geometric operation that scalar-reward RL cannot express and that yields the best overall system.

2 Related Work

Instruction-following TTS. Instruct-TTS asks a model to speak a target transcript in the manner described by a free-form natural-language instruction, such as a desired emotion, timbre, speaking rate, or dialect, and a good rendition must follow the instruction, stay intelligible, and sound natural at once. Early systems encode the style description with a dedicated prompt encoder and learn a mapping from the described attributes to acoustic style, as in PromptTTS (Guo et al., 2023) and InstructTTS (Yang et al., 2024). As large language models improved the understanding of free-form instructions, more recent systems such as FlexiVoice (Chen et al., 2026a), Mimo-Audio (Xiaomi, 2025), MOSS-VoiceGenerator (Huang et al., 2026), and Qwen3-TTS (Hu et al., 2026a) pair an LLM backbone with a speech decoder and follow much richer, open-vocabulary instructions.

Reinforcement learning for TTS. Reinforcement learning is increasingly used to post-train TTS models toward objectives that are not directly differentiable. Early efforts focus on autoregressive, LLM-based TTS in which speech is represented as discrete tokens, so the RL machinery closely follows that of text LLMs. Seed-TTS (Anastassiou et al., 2024) optimizes WER and speaker-similarity rewards with PPO, REINFORCE, and DPO; SpeechAlign (Zhang et al., 2024) aligns the synthetic distribution with a gold reference via DPO; DiffRO (Gao et al., 2025) adopts a token-level differen-

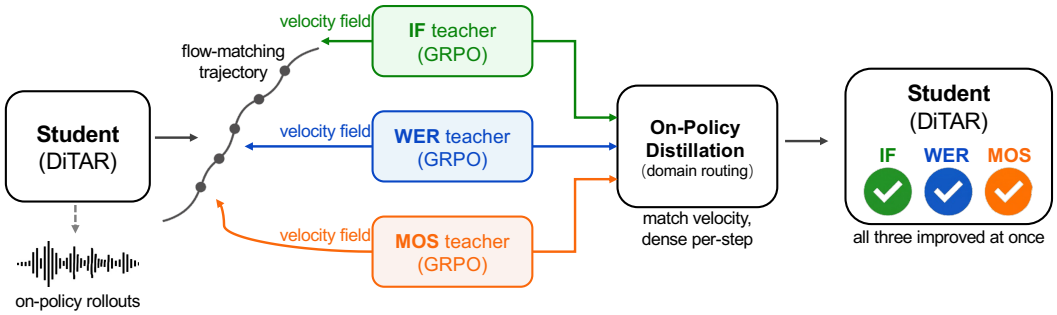

Figure 1: Overview of InstructTTS-OPD. Three single-objective teachers (IF, WER, MOS), each trained with GRPO, are evaluated on the student's own on-policy flow-matching rollouts and supply dense per-step velocity-field targets. Domain routing and conflict-aware fusion combine these targets so the student improves all three objectives at once, avoiding the scalar-reward seesaw.

tiable ASR reward; RRPO (Wang et al., 2026a) designs a more robust reward that alleviates reward hacking; and GRPO-TTS (Liu et al., 2026a) combines GRPO with a composite CER and NLL reward. A second line performs RL on flow-matching TTS: F5R-TTS (Sun et al., 2025) reformulates the deterministic flow into a probabilistic one to make policy gradients available, at the cost of retraining, whereas FlowTTS-GRPO (Wang et al., 2026b) turns the sampling ODE into an SDE and fine-tunes an open-source flow model directly with GRPO. Built on a DiTAR-style architecture, our work follows this second line and carries out RL and distillation in the continuous flow-matching space rather than over discrete tokens.

On-policy distillation. On-policy distillation (OPD) was first introduced for LLMs, where a student learns to match its teachers on the rollouts it generates itself, and has proven effective at combining several teachers into one student while easing the trade-off among them (Agarwal et al., 2024; Gu et al., 2023; Ma et al., 2026a; Xu et al., 2026a). The idea was later carried to image generation by Flow-OPD (Fang et al., 2026) and DiffusionOPD (Li et al., 2026c), which recast distillation on a flow-matching or diffusion model as matching the teacher velocity fields along the student's own rollouts; Flow-OPD, for instance, jointly optimizes GenEval, OCR, and aesthetic quality, and Qwen-Image-2.0-RL (Xu et al., 2026c) applies OPD to merge a generation and an editing model into one. We bring OPD to Instruct-TTS, optimizing instruction following, intelligibility, and quality together; the characteristics of speech introduce challenges the image setting does not face.

3 Method

3.1 Preliminaries: DiTAR backbone and flow matching

We build on a DiTAR-style architecture (Jia et al., 2025), which couples an autoregressive language model with a local flow-matching decoder. At each step the language model reads the text together with a summary of the audio generated so far and emits a condition vector c; a flow-matching decoder, a DiT block we refer to as the local decoder, turns Gaussian noise into the audio latents of the current frame under that condition; an aggregation encoder compresses the new frame back into a single vector that is fed to the language model as the next input, closing the autoregressive loop. A frozen vocoder finally maps the latents to a waveform. The language model is a 0.5B Qwen (Yang et al., 2025); together with the local decoder (193M) and the aggregation encoder (48M) the DiTAR model totals about 0.7B parameters (the vocoder is frozen). This design keeps generation autoregressive at the frame level while making each frame the output of a continuous flow-matching process, which is exactly where we apply reinforcement learning and distillation.

Flow matching (Lipman et al., 2022; Liu et al., 2022) learns a velocity field $v_\theta(x_t, t)$ that transports a Gaussian prior at $t{=}0$ to the data distribution at $t{=}1$. Given noise x_0 and a target latent x_1, one defines the straight-line path $x_t = (1-t)\,x_0 + t\,x_1$ with corresponding velocity $u_t = x_1 - x_0$, and trains v_θ to regress u_t. At inference, generation starts from noise and integrates the ordinary differential equation (ODE) $\mathrm{d}x = v_\theta(x_t, t)\,\mathrm{d}t$ from $t{=}0$ to $t{=}1$. Throughout, we write μ_θ for the

mean that the velocity field induces on one sampling step; it is the quantity our RL and distillation objectives act on.

3.2 From ODE to SDE: enabling on-policy rollouts

Deterministic ODE sampling maps a fixed noise sample to a single output and therefore offers no exploration, which reinforcement learning requires. We follow the practice of converting part of the sampling trajectory into a stochastic differential equation (SDE) (Wang et al., 2026b; Liu et al., 2026b): for a subset of the integration steps we add Gaussian noise around the deterministic mean, so that one step produces a sampled next state x', a deterministic mean $\mu_\theta(x_t, t)$, and the log-probability of the transition,

$$x' = \mu_\theta(x_t, t) + \sigma_t\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \log \pi_\theta(x' \mid x_t) = -\frac{\|x' - \mu_\theta(x_t, t)\|^2}{2\sigma_t^2} + \text{const.} \quad (1)$$

Sampling the same prompt several times now yields distinct rollouts, and each rollout records, at every SDE step, the state, condition, sampled output, and transition log-probability. The mean μ_θ carries gradients and is the shared optimization target of both routes described below: the group-relative policy gradient and the distillation loss.

3.3 Single-objective teachers via GRPO

We first train one expert per objective (IF, WER, and MOS) with group relative policy optimization (Shao et al., 2024). The three rewards are deliberately simple. The IF reward is a binary consistency verdict from a judge that decides whether an utterance matches its instruction, $r_{\mathrm{IF}} \in \{0, 1\}$ (a 3-vote majority); the WER reward is $r_{\mathrm{WER}} = 1 - \min(\mathrm{CER}, 1)$ from an ASR model; and the MOS reward is a no-reference quality score $r_{\mathrm{MOS}} \in [1, 5]$. For each prompt we draw a group of G on-policy rollouts, score each with the corresponding reward r, and standardize the scores within the group to obtain the advantage

$$A_i = \frac{r_i - \mathrm{mean}(\{r_j\}_{j=1}^G)}{\mathrm{std}(\{r_j\}_{j=1}^G)}, \quad (2)$$

so that only the relative quality within a group drives the update. The policy is then trained with the group-relative policy-gradient objective together with a KL regularizer to a frozen reference that prevents the base speech quality from drifting,

$$\mathcal{L}_{\mathrm{GRPO}} = -\,\mathbb{E}\big[A\,\log \pi_\theta\big] + \beta\,\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})\,. \quad (3)$$

This yields three single-objective experts, each strong on its own axis but weak on the others. Fusing them under a single scalar reward, however, requires collapsing the per-objective advantages into one number; the direction along which the objectives disagree is then lost, and optimization drifts toward the objective that is easiest to improve. This is the seesaw effect that motivates the distillation-based fusion below.

3.4 Multi-objective fusion via on-policy distillation

Rather than fusing the objectives through a scalar reward, we fuse the three experts directly in the velocity-field space with on-policy distillation (Fang et al., 2026; Li et al., 2026c), as illustrated in Figure 1. The student generates its own on-policy rollouts, and each teacher is evaluated at the same trajectory points that the student visits, so the supervision is delivered exactly where the student's own sampling takes it.

Because the student and every teacher share the same flow-matching transition kernel, the two single-step transition distributions are Gaussians with the same covariance, and their per-step Kullback–Leibler (KL) divergence collapses in closed form to a squared distance between their means,

$$\mathrm{KL}(\pi_{\mathrm{student}} \,\|\, \pi_{\mathrm{teacher}}) = \frac{\|\mu_{\mathrm{student}} - \mu_{\mathrm{teacher}}\|^2}{2\,\sigma_t^2\,\Delta t}. \quad (4)$$

The distillation signal is therefore a fully differentiable, zero-variance regression on the velocity-field mean, and it needs neither advantages nor a policy-gradient estimator.

The key question is how to combine several teachers. We route each rollout to a single main teacher by the domain of its training data: rollouts drawn from the instruction-following data are distilled against the IF teacher, and rollouts from the WER data against the WER teacher, with the overall mixing set by the data ratio. The reason to route the primary objectives to *disjoint* rollouts rather than have every rollout follow both teachers at once is the same conflict identified for the WER reward in Section 4.2: the WER data cannot contain dialect instructions, because a WER teacher scored by a Mandarin ASR would then push the model to abandon the requested dialect to lower its error rate, directly fighting instruction following. Keeping the IF and WER signals on separate rollouts prevents this tug-of-war on any single utterance. Quality, by contrast, has no domain of its own; it is a property every utterance should have. We therefore keep the MOS teacher out of the routing and add it as a light auxiliary term applied to all rollouts. The student minimizes

$$\mathcal{L}_{\mathrm{OPD}} = \left\|\mu_{\mathrm{student}} - \mathrm{sg}(\mu_{d(i)})\right\|^2 \;+\; \lambda_{\mathrm{MOS}} \left\|\mu_{\mathrm{student}} - \mathrm{sg}(\mu_{\mathrm{MOS}})\right\|^2, \tag{5}$$

where $d(i) \in \{\mathrm{IF}, \mathrm{WER}\}$ is the main teacher selected by the data domain of rollout i, $\mathrm{sg}(\cdot)$ is the stop-gradient operator that freezes the teachers as fixed targets, and λ_{MOS} is a small weight that makes quality a guardrail rather than a competing objective.

3.5 Conflict-aware velocity fusion

Adding the two targets in Eq. 5 treats the primary target and the quality target as two independent regressions whose gradients are simply summed, which is benign only when the two point in compatible directions. But quality and instruction following often disagree on a single utterance: naturalness pushes toward a smooth, evenly-paced delivery, whereas following an expressive instruction pushes toward crying, trembling, or dialectal speech that a quality predictor reads as degradation. On the frames where the two disagree, the summed gradient partially cancels, and raising λ_{MOS} to buy quality drags the primary objective down, exactly the seesaw we set out to avoid. We quantify how often this happens in Section 5.1.

We therefore fuse the two targets geometrically rather than additively. Writing the primary and quality update directions at a frame as

$$d_P = \mu_{d(i)} - \mu_{\mathrm{student}}, \qquad d_U = \mu_{\mathrm{MOS}} - \mu_{\mathrm{student}}, \tag{6}$$

we remove from d_U only the component that directly opposes d_P, keeping the aligned and orthogonal part,

$$d_U^{\mathrm{keep}} = d_U - \frac{\min(\langle d_U, d_P\rangle,\, 0)}{\|d_P\|^2}\, d_P, \qquad \mathcal{L}_{\mathrm{OPD}}^{\mathrm{conflict}} = \left\|\mu_{\mathrm{student}} - \mathrm{sg}\Big(\mu_{\mathrm{student}} + d_P + \lambda_{\mathrm{MOS}}\, d_U^{\mathrm{keep}}\Big)\right\|^2. \tag{7}$$

When the two directions form an obtuse angle ($\langle d_U, d_P\rangle < 0$) the opposing component is projected out, so $d_U^{\mathrm{keep}} \perp d_P$ and quality can no longer erode the primary objective; when they are aligned ($\langle d_U, d_P\rangle \geq 0$) the quality push is kept in full. Quality is thus only ever allowed to make orthogonal adjustments: it may refine the timbre or smoothness of a frame, but it can never flatten out an already-expressed sob or dialect. This borrows the idea of gradient surgery (Yu et al., 2020), but applies it to the per-frame velocity-field distillation targets of the flow model rather than to parameter gradients. It is an asymmetric, output-space geometric operation that a scalar-reward policy gradient cannot express, and that becomes available precisely because the distillation signal is a differentiable regression on μ. Because the target is stop-gradient, the loss reduces to a single mean-squared error and λ_{MOS} scales the surviving quality push linearly; InstructTTS-OPD uses this conflict-aware form as its fusion loss.

4 Experiments

4.1 Setup

Data and base model. All models start from a single supervised DiTAR checkpoint (the base) and are post-trained with RL and OPD on 30,000 instruction-following and 30,000 intelligibility (WER) utterances, a 1:1 ratio. Every instruction is in the ZH-DSD style, matching the base model, which has only seen ZH-DSD-style instruction data. We therefore evaluate primarily on the DSD

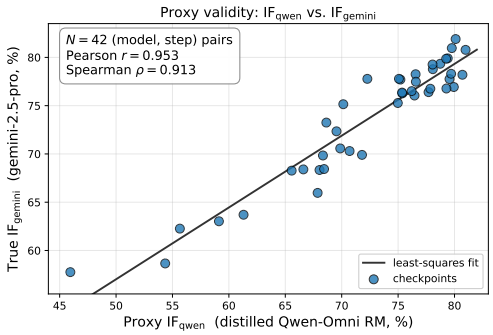

Figure 2: Distilled Qwen-Omni IF proxy vs. the Gemini judge, one point per (model, step) checkpoint scored on zh_DSD. x = proxy $\mathrm{IF}_{\mathrm{qwen}}$, y = Gemini IF.

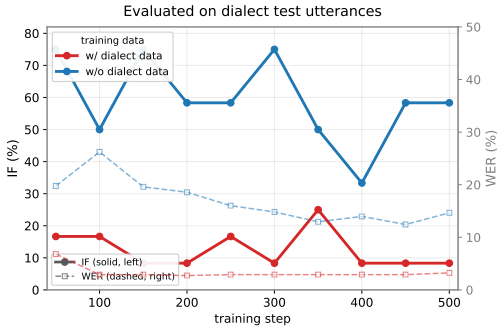

Figure 3: WER teacher's performance on the dialect (Cantonese) test subset across training, with vs. without dialect data. Solid lines are IF judged by Gemini on the left axis; dashed lines are WER on the right axis.

Table 1: ASR back-end for the WER reward: Corpus CER (%) on the full zh_DSD set (with dialect). Base = our supervised model; GT = human recordings.

Speech	Paraformer-zh	Qwen3-ASR
Base	3.51	3.17
GT	9.84	8.12

Table 2: Language bias of MOS predictors on matched clean human recordings (seed-tts-eval).

MOS predictor	zh	en	|gap| / range
UTMOS	2.774	3.491	**17.9%**
DNSMOS OVRL	3.183	3.071	2.8%
DNSMOS SIG	3.587	3.455	3.3%
DNSMOS BAK	3.854	3.820	0.8%

subset of InstructTTSEval (Huang et al., 2025) (zh_DSD, 1000 utterances), and also report the full six InstructTTSEval subsets to probe generalization.

Metrics. Each axis is optimized against a single chosen reward, but we report every axis with *both* its reward-side metric and a held-out metric that never shapes training, so that reward hacking would surface as a divergence between the two. Instruction following (IF), the primary axis, is the consistency rate of a judge that decides whether an utterance matches its instruction: a distilled Qwen2.5-Omni proxy ($\mathrm{IF}_{\mathrm{qwen}}$) for analysis and checkpoint selection, and Gemini as the held-out judge. Intelligibility is the character error rate (CER, ↓) from Paraformer-zh under the seed-tts-eval protocol, held out against Qwen3-ASR ($\mathrm{CER}_{\mathrm{q3}}$) and reported de-dialect. Quality is DNSMOS P.835 OVRL (↑), held out against UTMOS.

Training. Teachers are trained with GRPO and fused with OPD on 4 nodes × 8 GPUs for 350 steps. GRPO draws G=32 rollouts per prompt with a KL coefficient β=0.01; OPD uses 16 rollouts, needs no reward server, and sets the quality guardrail weight to λ_{MOS}=0.3. Both use a learning rate of 2×10^{-6}. Full hyperparameters are in Appendix A.

4.2 Reward Analysis

OPD fuses three single-objective teachers, so the quality of the fusion rests on having a strong teacher for each axis, which in turn depends on a sound reward. We therefore carefully analyze the reward choice along each of the three axes before training the teachers.

Instruction following: a distilled proxy for the Gemini judge. The standard way to score IF for Instruct-TTS is Gemini-as-a-judge (Yang et al., 2024; Chen et al., 2026b; Li et al., 2026b), which is reliable but far too expensive to query for the millions of rollouts an RL run consumes. We therefore distill a Qwen2.5-Omni-7B (Xu et al., 2025) reward model on Gemini judgments and ask whether it can serve as the IF reward in its place. Figure 2 scores 42 checkpoints spanning a wide IF range with both judges and plots one against the other: the points fall close to a straight line, and the two judgments are strongly correlated (Pearson r=0.95, Spearman ρ=0.91). The distilled model thus

Table 3: Main comparison on the zh_DSD subset of InstructTTSEval. Avg is the mean of IF, CER, and DNSMOS after min–max normalizing each to $[0, 1]$ with CER inverted. Per column, best in **bold**, second underlined.

System	IF ↑	CER ↓	CER_{q3} ↓	DNSMOS ↑	UTMOS ↑	Avg ↑
Ground-truth (human)	78.5	8.70	7.69	3.00	1.73	0.36
Base (supervised)	72.2	3.18	3.08	2.81	1.95	0.48
Teacher: IF	**82.1**	2.53	2.40	2.98	1.99	0.70
Teacher: WER	71.0	**1.94**	**2.00**	3.25	2.65	0.70
Teacher: MOS (quality)	55.0	2.60	2.51	**3.70**	**3.62**	0.63
GRPO multi-reward	77.0	2.34	2.27	3.25	2.48	0.75
InstructTTS-OPD	81.1	2.25	2.21	3.28	2.57	**0.82**
w/o conflict-aware fusion	79.4	2.29	2.23	3.31	2.65	0.80

ranks systems and checkpoints consistently with Gemini, so we use it as a low-cost reward model in place of Gemini throughout training and development.

Intelligibility: the WER teacher's training data must exclude dialect. A WER reward scored by a Mandarin ASR conflicts with instruction following whenever the instruction asks for a dialect, and Figure 3 makes this concrete. Trained *with* dialect data, the WER teacher drives its Mandarin-ASR WER down sharply, yet its IF stays near 10%: the model has simply stopped speaking the requested dialect and reads plain Mandarin that the ASR transcribes happily, so the WER gain is spurious, bought by disobeying the instruction. Trained *without* dialect data, it keeps the dialect and its IF is three to seven times higher, even though the Mandarin ASR reports a much higher WER on those correct utterances. We therefore remove dialect from the WER teacher's training data and report CER de-dialect throughout. We also tried a stronger ASR back-end for expressive and dialectal speech, but Table 1 shows Paraformer-zh (Gao et al., 2022) and Qwen3-ASR (Shi et al., 2026) nearly indistinguishable (0.1–1 pp CER) on both generated and human speech, so we keep Paraformer-zh.

Quality: UTMOS is language-biased, DNSMOS is not. A quality reward is only safe if it scores speech, not language. Table 2 scores matched clean human recordings in Chinese and English, whose true quality is comparable, and reads off the gap each predictor introduces between the two. UTMOS (Saeki et al., 2022) rates English 0.72 points higher than Chinese, 18% of its range; this bias reflects its almost-entirely-English training data rather than the audio, and as a reward it would push the model toward a foreign-accented reading of Chinese. DNSMOS P.835 (Reddy et al., 2022) shows almost no such gap ($\leq 3\%$ across all sub-scores), so we adopt DNSMOS OVRL as both the quality reward and the reported metric.

4.3 Main Results

With the three rewards fixed, we train one GRPO teacher per objective and then fuse them. As Table 3 shows, each teacher is strong on its own axis and visibly weak on the others: the IF teacher reaches 82.1 IF but only middling CER and quality, the WER teacher drives CER down to 1.94 at the cost of IF, and the MOS teacher tops DNSMOS while its IF collapses to 55.

We compare three fusion strategies under matched teachers and data. The scalar-reward baseline (*GRPO multi-reward*) still seesaws even with group-decoupled normalization (Liu et al., 2026d), improving the easier axes of quality and intelligibility while leaving the primary objective, IF, the lowest among all fused systems. On-policy distillation fares much better: our full method attains the highest IF of any fused system, close to the IF teacher's, together with the best CER while keeping DNSMOS strong. Removing conflict-aware fusion (*w/o conflict-aware fusion*) drops IF, since the quality target then conflicts with instruction following. Here Avg is the mean of the three optimized objectives after min–max normalizing each to $[0, 1]$ with CER inverted, and the held-out checks CER_{q3} and UTMOS are excluded from it.

To confirm these gains are genuine rather than reward overfitting, Table 3 also reports two metrics that never shape training: CER from a different ASR back-end (CER_{q3}) and UTMOS. Both preserve

Table 4: Instruction following on InstructTTSEval (Gemini judge, %, ↑), across all six subsets: {ZH, EN} × {APS, DSD, RP}. CV = CustomVoice, VD = VoiceDesign. Best among open systems in bold, second underlined.

Model	InstructTTSEval-ZH			InstructTTSEval-EN		
	APS (↑)	DSD (↑)	RP (↑)	APS (↑)	DSD (↑)	RP (↑)
Gemini-flash (Comanici et al., 2025)	88.2	90.9	77.3	92.3	93.8	80.1
Gemini-pro (Comanici et al., 2025)	89.0	90.1	75.5	87.6	86.0	67.2
GPT-4o-mini-tts (Hurst et al., 2024)	54.9	52.3	46.0	76.4	74.3	54.8
VoxInstruct (Zhou et al., 2024)	47.5	52.3	42.6	54.9	57.0	39.3
VoiceSculptor (Hu et al., 2026b)	75.7	64.7	61.5	–	–	–
Mimo-Audio-7B-Instruct (Xiaomi, 2025)	75.7	74.3	61.5	80.6	77.6	59.5
Qwen3TTS-25Hz-1.7B-CV (Hu et al., 2026a)	83.1	75.0	<u>63.0</u>	79.0	**82.8**	**69.3**
Qwen3TTS-12Hz-1.7B-CV (Hu et al., 2026a)	83.0	77.8	61.2	77.3	77.1	63.7
Qwen3TTS-12Hz-1.7B-VD (Hu et al., 2026a)	85.2	81.1	**65.1**	**82.9**	<u>82.4</u>	<u>68.4</u>
Base (Ours, before RL)	86.0	76.9	54.4	79.2	69.5	46.4
IF teacher (Ours)	**90.9**	**83.9**	58.7	<u>81.3</u>	75.6	48.8
InstructTTS-OPD (Ours)	<u>90.6</u>	<u>82.1</u>	56.2	80.8	73.4	47.2

the same ordering across systems as their reward-side counterparts, confirming that the gains are real. UTMOS values are uniformly low because it is biased against Chinese, so we read it only for the ranking it induces, which is exactly why DNSMOS is our quality reward and headline metric.

4.4 Results on InstructTTSEval

Although our model has only ever seen instructions in the ZH-DSD style, we want to know whether its instruction-following ability generalizes to unseen instruction forms. We therefore evaluate on all six subsets of InstructTTSEval, {ZH, EN} × {APS, DSD, RP}, scoring every system with the standard Gemini judge for a fair comparison; the three ZH subsets use the original instructions and the three EN subsets use their Chinese translations. Table 4 reports the results against both open and closed baselines. Post-training improves over the base on every subset, with the largest gains on the in-distribution DSD style and smaller but consistent gains on the harder APS and RP styles and on the out-of-distribution English side; among open systems our model is strongest or second on the Chinese subsets and competitive on English, while a gap to the closed Gemini systems remains on the RP style and on English. The improvements thus hold across languages and instruction styles well beyond the training distribution, showing that the fused model generalizes rather than overfitting the ZH-DSD form it was trained on.

5 Analysis

5.1 Per-frame conflict geometry

Conflict-aware fusion projects out the component of the quality push d_U that opposes the primary push d_P, so whether it matters depends on how often the two directions actually conflict. We measure this directly. Figure 4 tracks the per-frame angle θ between d_U and d_P over training: within about fifteen steps the median cosine similarity turns negative, aligned frames collapse to essentially zero, and conflicting frames climb to nearly all of them, with strongly-opposed frames still growing thereafter. Once the student becomes expressive, quality opposes the primary objective on almost every frame. This near-total conflict is exactly why the fusion rule matters: adding the quality push without projection would carry the opposing component almost everywhere and pull instruction following down, whereas the projection removes it frame by frame and keeps instruction following near the teacher's ceiling.

5.2 Trade-offs Analysis

The conflict above is about gradients; we now ask whether it surfaces in the end metrics, and find that the answer depends sharply on which pair of objectives we look at. We sweep two knobs and read off the frontiers in Figure 5, with full numbers in Appendix B. Sweeping the instruction:WER data

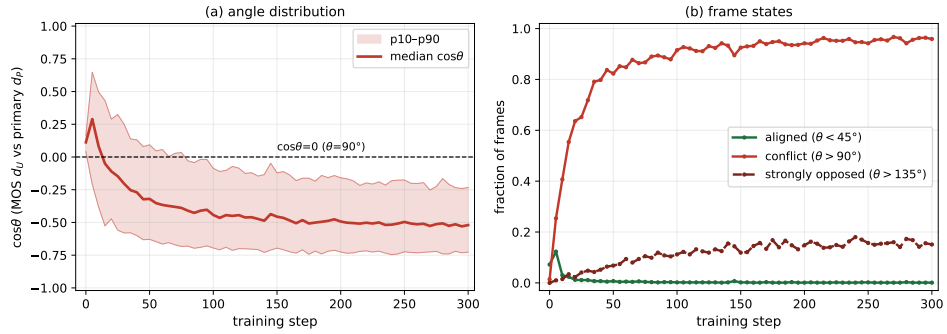

Figure 4: Per-frame geometry between the quality direction d_U and the primary direction d_P over training. **(a)** the cosine-similarity distribution, shown as the median with a 10 to 90% band; **(b)** the fractions of aligned ($\theta < 45°$), conflicting ($\theta > 90°$), and strongly-opposed ($\theta > 135°$) frames.

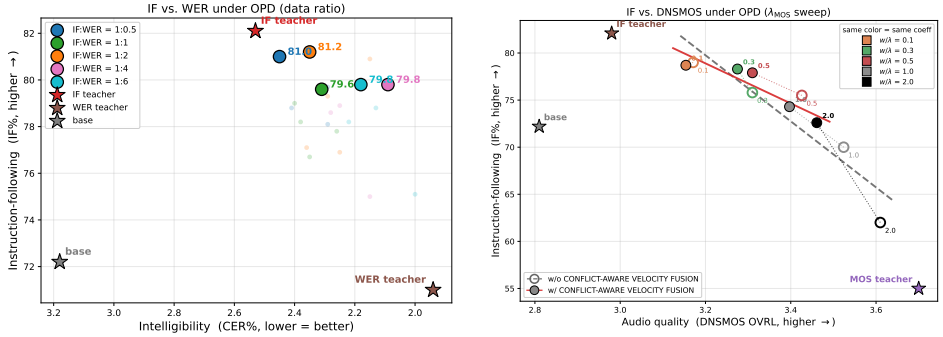

Figure 5: Trade-offs under OPD. **Left:** IF against WER while sweeping the instruction:WER data ratio; large dots mark the best-IF checkpoint per ratio, faint dots all checkpoints. **Right:** IF against DNSMOS while sweeping the quality weight λ_{MOS}, with (filled) and without (hollow) conflict-aware velocity fusion; same color shares a λ_{MOS}, stars mark the teachers and base.

ratio traces out an essentially flat IF–WER frontier: adding WER data drives CER toward its floor without costing instruction following, so the two improve together with no meaningful trade-off, as expected when they are routed to disjoint rollouts. A further ablation in Appendix C suggests that this routing, rather than the dense velocity supervision alone, is the main factor preserving instruction following. Sweeping the quality weight λ_{MOS} tells the opposite story. Without conflict-aware fusion, instruction following falls steadily as the quality pressure rises, the metric-level shadow of the near-total per-frame conflict; conflict-aware velocity fusion flattens this trade-off, keeping instruction following far higher at matched weights because the projection stops the quality push from eroding it frame by frame.

6 Conclusion

We studied how to post-train an instruction-following TTS model on its three competing objectives without the seesaw a single scalar reward induces, and two lessons emerged. First, the reward for each axis matters as much as the fusion algorithm: a distilled Qwen-Omni proxy tracks the costly Gemini judge, the WER reward must be dialect-free, and DNSMOS replaces the language-biased UTMOS. Second, fusing single-objective teachers through on-policy distillation improves all three axes together where a scalar reward cannot: domain routing lets instruction following and intelligibility rise together, while quality, which conflicts with instruction following, is fused geometrically rather than additively. This per-frame geometric fusion is available only because distillation in continuous flow-matching space is a differentiable regression on the velocity field, and we hope it offers a practical recipe for multi-objective alignment of expressive speech generators.

AI USE STATEMENT

We used generative AI tools to polish the writing of this paper, improving its grammar and readability.

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 *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Dekun Chen, Xueyao Zhang, Yuancheng Wang, Kenan Dai, Li Ma, and Zhizheng Wu. Flexivoice: Enabling flexible style control in zero-shot tts with natural language instructions. *arXiv preprint arXiv:2601.04656*, 2026a.

Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, et al. Mint-bench: A comprehensive multilingual benchmark for instruction-following text-to-speech. *arXiv preprint arXiv:2604.17958*, 2026b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Changfeng Gao, Zhihao Du, and Shiliang Zhang. Differentiable reward optimization for llm based tts system. *arXiv preprint arXiv:2507.05911*, 2025.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*, 2022.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026a.

Jingbin Hu, Huakang Chen, Linhan Ma, Dake Guo, Qirui Zhan, Wenhao Li, Haoyu Zhang, Kangxiang Xia, Ziyu Zhang, Wenjie Tian, et al. Voicesculptor: Your voice, designed by you. *arXiv preprint arXiv:2601.10629*, 2026b.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructtts eval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025.

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, Jie Zhu, Yuqian Zhang, Yiwei Zhao, Chenchen Yang, Zhaoye Fei, et al. Moss-voicegenerator: Create realistic voices with natural language descriptions. *arXiv preprint arXiv:2603.28086*, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

Haitao Li, Chunxiang Jin, Chenglin Li, Wenhao Guan, Zhengxing Huang, and Xie Chen. Restyle-tts: Relative and continuous style control for zero-shot speech synthesis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9257–9269, 2026a.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026b.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Chang Liu, Ya-Jun Hu, Ying-Ying Gao, Shi-Lei Zhang, and Zhen-Hua Ling. Group relative policy optimization for text-to-speech with large language models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16132–16136. IEEE, 2026a.

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. *Advances in neural information processing systems*, 38:40783–40818, 2026b.

Qingyu Liu, Rixi Xu, Yushen Chen, Zhikang Niu, Haitao Li, Pengcheng Zhu, Bowen Zhang, Jian Zhao, Yunting Yang, Qinyuan Cheng, et al. Cross-lingual f5-tts 2: A simplified framework for language-agnostic voice cloning. *arXiv preprint arXiv:2609.15184*, 2026c.

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*, 2026d.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026a.

Ziyang Ma, Zhikang Niu, Wenming Tu, Tianrui Wang, Ruiqi Yan, Junxi Liu, Yanru Huo, Nickk Huang, Yang Liu, Qicong Xie, et al. Auk technical report: An open-source foundational model for speech generation and editing. *arXiv preprint arXiv:2609.08936*, 2026b.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 886–890. IEEE, 2022.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025.

Cong Wang, Changfeng Gao, Yang Xiang, Zhihao Du, Keyu An, Han Zhao, Qian Chen, Xiangang Li, Yingming Gao, and Ya Li. Rrpo: Robust reward policy optimization for llm-based emotional tts. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16497–16501. IEEE, 2026a.

Haoxu Wang, Biao Tian, Weiqin Li, Xiang Lv, Han Zhao, and Xiangang Li. Flowtts-grpo: Online reinforcement learning with multi-objective reward optimization for flow-matching based text-to-speech. *arXiv preprint arXiv:2606.23190*, 2026b.

LCT Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR, abs/2512.23808*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Rixi Xu, Qingyu Liu, Haitao Li, Yushen Chen, Zhikang Niu, Yunting Yang, Jian Zhao, Ke Li, Berrak Sisman, Qinyuan Cheng, et al. X-voice: Enabling everyone to speak 30 languages via zero-shot cross-lingual voice cloning. *arXiv preprint arXiv:2605.05611*, 2026b.

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, et al. Qwen-image-2.0-rl technical report. *arXiv preprint arXiv:2606.27608*, 2026c.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Guanrou Yang, Tian Tan, Qian Chen, Ziyang Ma, Yakun Song, Zhikang Niu, Qi Chen, Wenming Tu, Haitao Li, Shan Yang, et al. Grow: Group-relative advantage-weighted on-policy reinforcement learning of autoregressive-diffusion text-to-speech model. *arXiv preprint arXiv:2608.03215*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechalign: Aligning speech generation to human preferences. *Advances in Neural Information Processing Systems*, 37:50343–50360, 2024.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 554–563, 2024.

A HYPERPARAMETERS

Table 5 lists the training hyperparameters for the GRPO teachers and for OPD fusion. All runs start from the same supervised base checkpoint and train on 4 nodes $\times$ 8 GPUs. The flow-matching decoder uses a patch size of 5 latents per frame with a 10-step ODE, of which a contiguous 3-step window is converted to an SDE for on-policy rollouts, under classifier-free guidance α=2.0.

Table 5: Training hyperparameters. GRPO trains one single-objective teacher each (IF, WER, MOS); OPD fuses them.

	GRPO (per teacher)	OPD fusion
Rollouts per prompt G	32	16
KL coefficient β	0.01	0
Learning rate	2×10^{-6}	2×10^{-6}
Optimizer	Adam $(0.9, 0.95)$	Adam $(0.9, 0.95)$
Micro-batch / grad-accum	4 / 4	4 / 4
Checkpoint interval (steps)	50	50
Reward server required	yes	no
Fusion mode	—	domain routing
MOS guardrail λ_{MOS}	—	0.3
Shared flow-matching / SDE settings		
Latents per frame (patch size)	5	
ODE steps / SDE window	10 / 3	
CFG strength α	2.0	

B TRADE-OFF SWEEPS: FULL NUMBERS

This appendix gives the full numbers behind the two trade-off plots of Figure 5, with the single-objective teachers as the two extremes of each sweep. Table 6 lists the instruction:WER data-ratio sweep, reporting the best-IF checkpoint per ratio. Table 7 lists the quality-weight λ_{MOS} sweep, reporting both fusion rules, without and with conflict-aware velocity fusion, at a matched checkpoint (step 200) so that only the fusion rule differs between the two halves. All runs use the final three teachers and are evaluated on zh_DSD (n=1000), with IF from the Qwen2.5-Omni RM (%, ↑), CER from Paraformer-zh (de-dialect, %, ↓), and DNSMOS P.835 OVRL (↑).

Table 6: Instruction:WER data-ratio sweep (λ_{MOS}=0.1 fixed), best-IF checkpoint per ratio.

IF:WER	IF ↑	CER ↓	DNSMOS ↑
IF teacher	82.1	2.53	2.98
1:0.5	81.0	2.45	3.17
1:1	79.6	2.31	3.18
1:2	81.2	2.35	3.19
1:4	79.8	2.09	3.23
1:6	79.8	2.18	3.21
WER teacher	71.0	1.94	3.25

C ABLATION: IS DOMAIN ROUTING NECESSARY?

The fusion in Section 3.4 routes each rollout to a single main teacher by its data domain. A natural alternative is to distill every rollout against all teachers at once with a weighted-sum loss, the velocity-field analog of scalar-reward fusion. We compare the two under identical teachers, data, and budget (Figure 6). Routing keeps IF high throughout training, climbing toward the IF teacher, whereas weighted co-distillation stays several points lower and drifts downward from the first checkpoints, so the two IF trajectories never overlap. Both keep CER and DNSMOS within target, so the gap is concentrated on IF, the primary objective. The mechanism is the same intra-sample conflict as in the scalar case: when a single utterance is pulled toward instruction following, intelligibility, and quality at once, the competing velocities average into a bland compromise that caps IF. Routing

Table 7: Quality-weight λ_{MOS} sweep (IF:WER=1:1 fixed), without vs. with conflict-aware velocity fusion at a matched checkpoint (step 200). Everything but the fusion rule is held fixed, and teachers are the two extremes of the sweep.

	w/o conflict-aware			w/ conflict-aware		
λ_{MOS}	IF ↑	CER ↓	DNSMOS ↑	IF ↑	CER ↓	DNSMOS ↑
IF teacher	82.1	2.53	2.98	82.1	2.53	2.98
0.1	79.0	2.40	3.17	78.7	2.24	3.15
0.3	75.8	2.33	3.31	78.3	2.23	3.28
0.5	75.5	2.28	3.43	77.9	2.29	3.31
1.0	70.0	2.32	3.52	74.3	2.22	3.40
2.0	62.0	2.40	3.61	72.6	2.29	3.46
MOS teacher	55.0	2.60	3.70	55.0	2.60	3.70

sidesteps this by sending the competing primary objectives to disjoint rollouts, which suggests that it is the routing, not merely the dense velocity supervision, that makes the fusion work.

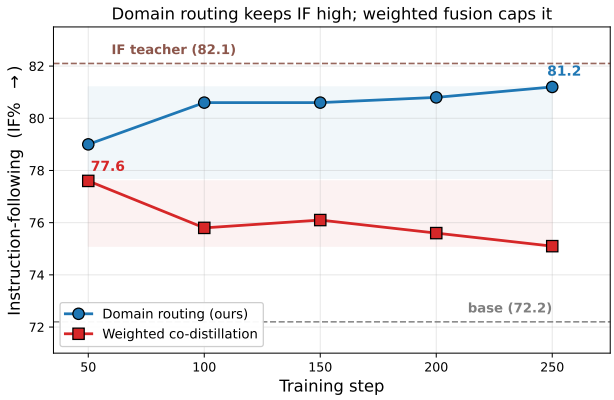

Figure 6: Domain routing vs. weighted co-distillation: IF (Qwen-Omni RM) at every checkpoint through convergence, under identical teachers and 1:1 data.

D HUMAN PREFERENCE EVALUATION

To check that the automatic-metric gains translate into perceived quality, we run a blind A/B listening test comparing the final fused model after RL against the supervised base checkpoint before RL. Raters hear the two utterances for each prompt in randomized order and choose which one better follows the instruction while sounding natural, or mark a tie. Aggregating over all trials, the after-RL model wins 35.0% of comparisons, loses 28.3%, and ties 36.7% (Figure 7), so it is preferred over the base in 6.7 points more comparisons than it loses, confirming that the objective-metric improvements carry over to human judgement.

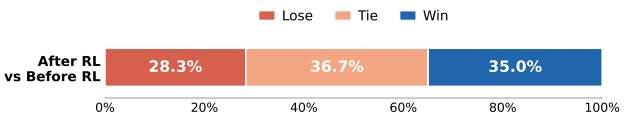

Figure 7: Blind A/B human preference, after RL vs. before RL. Bars show the share of comparisons the after-RL model loses, ties, and wins.