ITERCAP: A SELF-ANNOTATING CAPTIONING FRAMEWORK FOR INSTRUCTION-FOLLOWING TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-following text-to-speech (instruct-TTS) learns acoustic controllability from captions, making their quality and coverage fundamental to what the model can learn: an attribute that a caption fails to specify or ground receives little supervision, regardless of the synthesizer's capacity. Existing methods either take captions as given, whether human-written, generated zero-shot by an external audio-language model, or produced by a handcrafted pipeline, or train a caption model on such captions once through supervised fine-tuning or a single round of reinforcement learning. In all cases, the supervision originates outside the caption model and is fixed before the model trains on it. We instead ask whether a caption model can annotate its own supervision, formulating instruct-TTS captioning as a self-annotating problem that exploits an asymmetry: verifying the attributes asserted by a caption is easier than generating a correct open-ended caption. IterCap closes this loop through three operations: selection verifies generated captions with lightweight acoustic-consistency checks and independent caption models; expansion adds examples of underrepresented delivery conditions; and optimization converts field-level correctness into a reference-anchored, verifiable objective through a multiple-choice exam and Group Sequence Policy Optimization (GSPO). On a 325-clip human-annotated test set spanning 12 controllable dimensions, IterCap raises macro-averaged per-dimension caption accuracy to 83.4%, compared with 75.3% for the strongest open-source caption model. This improvement also transfers to the downstream system: with the 300K-sample corpus and Qwen3-TTS backbone held fixed, IterCap captions raise instruction-following accuracy to 88.3%, compared with 84.5% using the corpus's native captions and 82.9% using the zero-shot caption model from which IterCap is initialized.

1 INTRODUCTION

Instruction-following text-to-speech (instruct-TTS) does not learn controllability directly from audio; it learns it through language supervision. Captions are therefore not merely descriptions of speech but the supervision through which acoustic controllability becomes learnable: unlike reference-based synthesis, which copies the desired acoustic properties from a prompt recording, instruct-TTS has them specified only in language at training time, so the controllability it can acquire is bounded by what its training captions specify and ground (Hu et al., 2026; Li et al., 2026b). An attribute that a caption leaves unspecified or ungrounded receives no reliable supervision, however much capacity the downstream model has. The supervision itself therefore becomes an object to improve, not merely a fixed input to the model that consumes it.

Existing open-source audio caption models address this gap only partially. Their captions are obtained in one of several ways, each with a characteristic shortfall. Captions produced zero-shot by a general-purpose audio-language model (LALM) (Ma et al., 2026; Zhou et al., 2026b; Xu et al., 2025) cover the dimensions an instruct-TTS system must control, such as timbre, emotion, style, and persona, only sparsely and inconsistently; human-annotated sets are small and costly; generation with stronger external models is expensive; captions built by programmatic construction cover only the attributes their rules were written to extract (Kumar et al., 2026; Nie et al., 2026; Chin et al., 2026); and reinforcement learning improves caption quality only against an external reward

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.

(Ghosh et al., 2026b; Chen et al., 2026; Tang et al., 2025). The strongest instruct-TTS systems rely on captions from proprietary pipelines whose provenance is undisclosed (Hu et al., 2026; Li et al., 2026b). Where a dedicated caption model is trained, most work improves it by post-training on captions from one of these sources (Gong et al., 2024; Deshmukh et al., 2025; Xu et al., 2026a), so the signal it optimizes against is supplied from outside the loop. All of these are static in the same sense: each is a one-shot strategy, designed in advance, that leaves the supervision itself unchanged once it has been produced. The field is thus split: the captions capable of training a strong instruct-TTS system are not publicly obtainable, and the publicly obtainable ones fall short of what such a system requires.

A key asymmetry motivates our approach: judging whether a caption is correct is considerably easier than producing one. Generation requires a model to commit to a definite description of every controllable attribute at once, whereas verification only has to test the attributes already asserted against cheap and complementary evidence, such as whether the audio realizes them or whether independent caption models concur. Self-generated captions can therefore serve as training supervision once they are selectively verified, and such a loop needs no perfect generator, only a verifier that is easier to build than the generator it supervises. We call the loop self-annotating in a specific sense: the model's own captions are the only scalable source of textual supervision, and the external components are judges rather than teachers, deciding which of the model's own annotations are retained, while an acoustic editor constructs new inputs to expand support. Human annotation is never applied at corpus scale: it enters only as the evaluation sets and as the fixed answers of the verification exam, both of bounded size, both built once.

Closing the loop, however, is not a matter of re-annotating and retraining. A naive loop that does only that fails in three distinct ways, none a variant of the others. First, it reinforces its own incorrect annotations, since at corpus scale no human verification is available to catch them. Second, it cannot acquire capabilities absent from its current support, because filtering merely reweights what the model already produces. Third, it supplies no precise learning signal for individual fields: a verdict on a field is obtainable, but not a label reliable enough to train against on real samples. The three failures are a wrong signal, a missing one, and an imprecise one, and each calls for a different remedy.

We propose IterCap, a closed-loop supervision framework for instruct-TTS captioning whose three operations select what is correct, create what is missing, and optimize what is controllable. Our contributions are threefold.

- **Self-annotating supervision.** We formulate instruct-TTS captioning as a self-annotating supervision problem, in which the caption model generates its own annotations, external verification decides which of them survive, and the model retrains on the survivors, so its supervision originates in its own output rather than in a fixed external caption source. The formulation identifies three requirements for scalable self-improvement: supervision precision, support coverage, and field-level correctness.
- **Closed-loop supervision with capability expansion.** We develop IterCap, which combines rejection-based selection with controlled acoustic intervention. Selection retains only the captions that pass complementary acoustic-consistency and cross-model verification, while controlled editing introduces delivery conditions the natural training distribution does not contain, so the loop improves both the precision and the support of its supervision without per-round human annotation. With the corpus and the downstream backbone held fixed, the resulting captions raise downstream instruction following from 84.5 to 88.3.
- **Reference-anchored verifiable optimization.** We convert the field-level correctness of a free-form natural-language caption into a fixed multiple-choice examination built from human-annotated references, and optimize the resulting exact-match reward with GSPO, a signal that is verifiable, needs no judge at rollout time, and raises the caption model from 82.9 to 83.4 on macro accuracy.

We then re-annotate a public corpus with the resulting model and train Qwen3-TTS on it. IterCap reaches 83.4 on the macro average over the 12 caption dimensions, against 75.3 for the strongest open-source caption model, and the gain reaches the downstream system: with the corpus, the backbone, and the training recipe held fixed, IterCap captions raise instruction following from 84.5 to 88.3, against 82.9 for the same audio-language model used zero-shot as the caption source.

2 Related work

The captions used to train controllable and instruct-TTS systems are, in dominant practice, produced by prompting a general-purpose audio-language model and optionally filtering its output. Seven provenance regimes recur: human-written captions (Ghosh et al., 2024; Wu et al., 2026b), zero-shot LALM generation (Ma et al., 2026; Wu et al., 2026a; Zhou et al., 2026b), generation checked by a human or a signal-level verifier (Tang et al., 2024; Gong et al., 2024; Chen et al., 2025), programmatic construction (Kumar et al., 2026; Nie et al., 2026; Chin et al., 2026), reinforcement learning against an external reward (Xing et al., 2026; Zhang et al., 2025; Ghosh et al., 2026b), self-bootstrapping re-annotation (Yang et al., 2026a; Tang et al., 2025), and undisclosed pipelines (Hu et al., 2026; Li et al., 2026b); Appendix B tabulates them paper by paper. What unites the regimes is that they are one-shot or externally supervised: a caption model either annotates once or is trained against a fixed external signal, and none lets the caption model improve its own supervision across successive rounds of verification and re-annotation. The closest prior work is MOSS-Audio (Yang et al., 2026a), whose speech-caption branch bootstraps an annotator from an internal captioning model through an offline two-stage procedure; its production annotator remains a mixture of a small and a frontier model, its schema acts as a post-hoc filter on generated captions, and its RL stage targets reasoning. video-SALMONN 2 (Tang et al., 2025) re-annotates existing data with its own trained model. Our loop differs in what a round changes: each annotation is verified against the signal it describes, the delivery conditions available to the model are extended beyond those the pool contains, and correctness is optimized field by field, so a round alters the precision and the support of the supervision and not only its wording.

Three further lines of work supply components that IterCap recombines; Appendix A discusses them in full. Rejection sampling fine-tuning (Yuan et al., 2023), and the self-bootstrapping loops it descends from (Zelikman et al., 2022; Gulcehre et al., 2023; Singh et al., 2023), contribute the accept-and-retrain mechanism together with the limitation that a selective operator cannot introduce behaviors absent from the output distribution; we circumvent it by controlled editing and guard the verdict with a two-stage verifier, since a model's own reward judgments are unreliable (Zhou et al., 2025). Typed caption schemas (Wu et al., 2026b; Kumar et al., 2026) contribute the field decomposition, which here is a training-time constraint that makes a caption machine-checkable rather than an evaluation protocol over professional annotations; unlike interfaces that condition synthesis on attribute tags or embeddings (Huang et al., 2026; Liao et al., 2026; Zhou et al., 2026a), every field holds a free-form phrase, so the schema fixes what a caption must describe without bounding what it may say. Multiple-choice (MCQ) reformulation of open-ended captioning into a verifiable reward (Xing et al., 2026; Yang et al., 2026c; Zhang et al., 2025) and GSPO (Zheng et al., 2025) contribute the optimization stage, which we anchor in fixed exam answers rather than in a learned judge, so that the reward neither drifts at rollout time nor inherits the errors of a reward model. Prior work thus supplies the individual ingredients of a supervision loop but not the object they would act on: each is applied to a fixed annotation source, so supervision remains an input to the caption model rather than something the model improves.

3 Method

3.1 Overview

Figure 1 summarizes the framework. Given a clip x, the caption model π_θ emits a structured caption $c \sim \pi_\theta(\cdot \mid x)$ as a set of typed fields covering timbre, emotion, and style. IterCap decomposes the problem into three operations: selection, a rejection-sampling flywheel keeping only the annotations that clear both verification stages; expansion, controlled editing that supplies delivery conditions absent from the corpus and from the model's own output support, together with the field value each edit sets; and optimization, MCQ-scored GSPO turning per-field verification into a reference-anchored exam. A held-out check after each round decides whether the loop moves on or the round is retrained with constructed audio added. The resulting model re-annotates the filtered subset of the public ISC-SLP 2026 CoT-TTS corpus (Xue et al., 2026), on which Qwen3-TTS is trained (Appendix C).

The three operations are not separable contributions; each depends on the other two. Selection keeps the other two sound: the constructed clips are captioned by the model itself and the exam sharpens field choices within what the model already writes, so without an acceptance test both

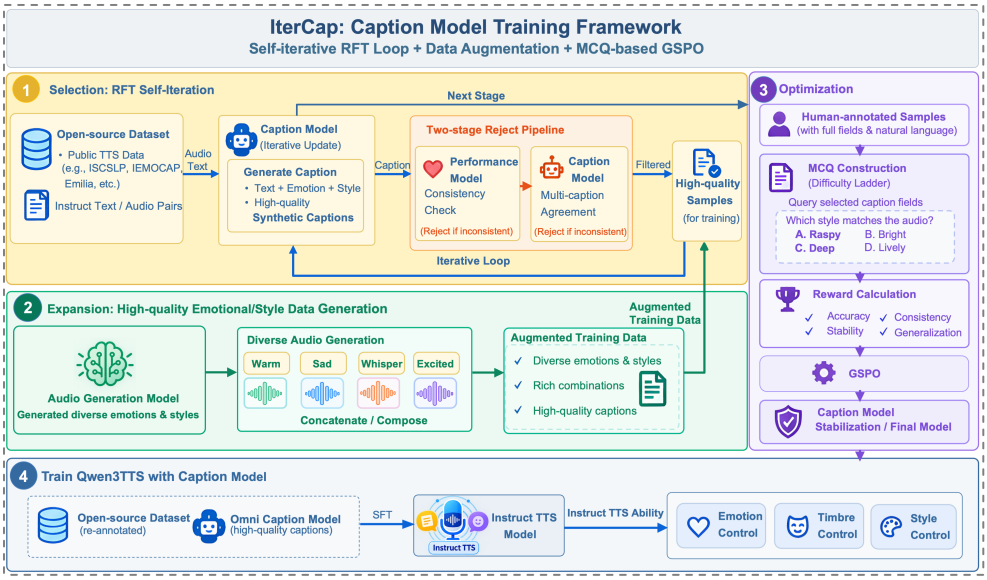

Figure 1: Overview of IterCap. A dedicated caption model is trained through three operations: selection, which retains the captions clearing both verification stages; expansion, which supplies constructed audio for delivery conditions the corpus lacks; and optimization, which applies MCQ-scored GSPO.

would propagate the errors they inherit. Expansion keeps them from converging inside a fixed support: an acceptance test can only reweight the material it is given, and the exam can only score dimensions the model is in a position to express. Optimization supplies what selection cannot: acceptance decides a candidate as a whole, so a wrong field inside an accepted caption is trained on with the rest. Removing an operation therefore withdraws a property the remaining two rely on rather than a third of the effect. Expansion is invoked for the first round and thereafter only when selection alone has stopped improving the model on the development set, which is the condition the round-wise check tests for. This dependency is what addresses the three structural failures of Section 1.

3.2 Selection: The Self-Annotating Supervision Flywheel

The first operation, selection, addresses the precision of the supervision and implements the core self-annotating loop of IterCap. At iteration t, the current caption model π_{θ_t} captions the unlabeled training pool, and every candidate passes through two independent verification stages; those that clear both stages, and whose clips the accumulated annotation set does not already cover, are retained as training data. The model is then trained on the expanded set, producing $\pi_{\theta_{t+1}}$, which is scored on a small held-out development set before the next round starts. When the score rises the loop moves on to the next round on natural audio alone, and when it does not the round itself is retrained, its training set augmented with constructed audio targeting the dimensions that failed to improve. Beyond the first round, every injection of constructed audio is therefore preceded by a round of selection alone that failed to improve. The first round has no earlier score to compare against, so it trains on its accepted annotations together with a constructed batch from the outset, the initial model's annotations covering the dimensions too unevenly for selection alone to broaden them.

The reference descriptions the verification models produce are used only for comparison and never enter the training set, so a candidate is accepted or rejected, never replaced. A verifier therefore need not be a stronger captioner than the model it screens; the stage exploits the asymmetry noted in Section 1: deciding whether an asserted attribute holds is far easier than composing a correct caption. A verdict does carry information about the verifier's preferences, but it cannot add, only

withhold: it can suppress a description the model produced, never introduce one it would not have, so the annotations the loop can reach stay inside the model's own output support.

Formally, let θ_t be the parameters at round t and $\mathcal{D}_t$ the set of pairs (x, c) with $c \sim \pi_{\theta_t}(\cdot \mid x)$ over the training pool. Each pair is kept only if it clears two independent stages: an acoustic stage, in which a lightweight scorer checks whether the performance realized in the audio matches what the caption asserts, and a cross-model stage, in which one or more independent caption models annotate the same clip under the same schema and a text-only judge scores their agreement with the candidate. The acceptance predicate conjoins them:

$$a(x, c) \;=\; a_{\text{acoustic}}(x, c) \;\wedge\; a_{\text{cross}}(x, c), \tag{1}$$

where $a_{\text{acoustic}}(x, c)$ holds when the scorer clears the caption on every axis it asserts, and $a_{\text{cross}}(x, c)$ when the candidate agrees with every independent model's own annotation of the clip at or above the round's cut. Neither stage is tied to a particular model: any sufficiently capable emotion/style scoring model may serve as the acoustic-consistency scorer, and any number of open- or closed-source caption models as the cross-model stage.

A two-stage verifier. A caption model cannot reliably detect errors in its own output: the failure that matters is not an absurd caption but a fluent yet ungrounded one, naming attributes the audio does not carry. The two stages therefore differ in granularity rather than checking the same thing twice. The acoustic stage is a cheap coarse screen that tests each asserted field against acoustic statistics and rejects only when the audio plainly contradicts it in either direction, an expressive claim over a flat delivery or a flat claim over an expressive one; the cross-model stage is a finer semantic check that catches the grounding errors the screen lets through. Acceptance requires both.

The two stages in our experiments. The framework fixes what each stage must decide, not which model decides it; the instances we use to validate it are replaceable. For the acoustic stage we use DeEAR (Lin et al., 2025), the performance model of Figure 1 and far lighter than the caption model; for the cross-model stage we use two open-source caption models, MOSS-Audio-8B-Instruct (Yang et al., 2026a) and Audio Flamingo 3 (Ghosh et al., 2026a), each annotating the same clip independently under our schema, with Qwen2.5-7B-Instruct (Yang et al., 2024) scoring the candidate's agreement with each. Appendix F gives the per-axis thresholds, the scoring protocol, the cost ordering that the conjunction permits, and their validation.

Round-wise training and exclusion. Each round starts from the base weights and trains on every annotation accepted so far, rather than on that round's own additions alone: each increment is small relative to the corpus, so the accumulated set keeps the data mix controllable and avoids the forgetting that fitting increments in isolation would induce. The exclusion rule, which governs what a round adds rather than what it trains on, is what distinguishes this loop from repeated rejection sampling fine-tuning (RFT); without it, each round would re-accept the clips it had already covered, sharpening a fixed subset instead of extending its coverage. Folding the exclusion into the acceptance indicator, $\tilde{a}(x, c) = a(x, c) \cdot \mathbf{1}\big[x \notin \mathcal{X}_{<t}\big]$ with $\mathcal{X}_{<t}$ the clips annotated in earlier rounds, the pairs round t adds are $\mathcal{A}_t = \{\, (x, c) \in \mathcal{D}_t : \tilde{a}(x, c) = 1 \,\}$ and its training set is the cumulative union $\mathcal{C}_t = \bigcup_{t' \le t} \mathcal{A}_{t'}$, on which the next model is obtained by maximum likelihood from θ_0:

$$\theta_{t+1} \;=\; \arg\min_{\theta'} \; -\frac{1}{|\mathcal{C}_t|} \sum_{(x,\, c) \,\in\, \mathcal{C}_t} \log \pi_{\theta'}(c \mid x), \tag{2}$$

where the minimization starts from $\theta' = \theta_0$ and each pair (x, c) in $\mathcal{C}_t$ is a clip from the training pool together with the annotation some round's model produced for it and both stages accepted.

3.3 Expansion: controlled acoustic intervention

The second operation, expansion, addresses the support of the supervision. The obstacle is distributional, and this is where expansion differs from selection in kind: rejection sampling fine-tuning conditions only on the acceptance event, so it acts on the caption model's output distribution $\pi_\theta(\cdot \mid x)$ as a reweighting, $\pi_\theta \to \pi_\theta \mid \text{accepted}$, which redistributes mass within the support but cannot enlarge it. A delivery condition to which π_θ assigns probability zero stays at zero however many rounds are run, so the limitation is one of support rather than of scale. Natural corpora make the gap acute:

delivery style is coupled with speaker identity, each speaker contributes a narrow slice of the emotion/style space, and many pairs are unattested. A selective flywheel can only amplify the coupling it starts with, so the rare combinations a user is most likely to request are exactly the ones it never sees.

Controlled editing. The property we exploit is that the operation is a controlled edit rather than generation from scratch: given a source clip x and a target expression z, the editor produces a variant $x' = E(x, z)$ with speaker(x') = speaker(x), content(x') = content(x) and style$(x') \neq$ style(x), re-rendering the same voice and words under a different emotion or style. This is intervention rather than augmentation. An observation records only what one speaker happened to deliver, whereas an intervention varies the delivery with speaker and content held fixed, so the caption model learns what a style label refers to rather than which voice it co-occurs with; we call the supervision it supplies *counterfactual supervision*.

What the intervention provides. The construction has two properties the flywheel cannot obtain from natural audio. First, one source utterance can be re-rendered into many emotion/style variants, and the variants concatenated into composite passages, enumerating pairs the flywheel would never see. Second, we set the target expression ourselves, so its value is known without asking the model. Selection cannot supply this: the flywheel keeps a caption only when the model already describes the clip correctly, so a dimension the model gets wrong contributes nothing and stays wrong, however many real recordings of it the corpus holds. An edit fixes that value at construction time.

Where the edited audio enters the loop. The resulting audio is captioned by the caption model and mixed into the training pool of the first round and of the rounds the development-set check sends back for retraining, so the injection is an external capability source rather than a caption source: it changes what the model can hear, and the caption's text remains the model's own apart from the single field the edit fixes. Constructed clips differ from natural ones in one respect. The target expression z is chosen before the edit is rendered, so the value of the field it sets is known at construction time: the model captions the clip as usual, that one field is then written to its constructed value, and every other field stays as the model wrote it and is verified as on natural audio. Expansion can therefore extend the support rather than resample it: a delivery the model cannot yet describe still yields supervision on the field that names it. In our experiments the editor E is Step-Audio-EditX (Yan et al., 2025), paired with CosyVoice 3 (Du et al., 2025) for target expressions no source clip carries. Neither is required by the methodology, which asks only for a system that places an utterance under a chosen delivery while holding timbre and transcript fixed.

3.4 Optimization: MCQ-scored GSPO

The third operation, optimization, addresses correctness field by field. The flywheel's supervised fine-tuning improves the precision of the supervision but is the wrong instrument for field-level correctness. Its objective scores surface form, not truth value: two captions asserting the same attribute in different words incur different losses, while two differing only in whether one field is right are nearly indistinguishable, the disputed field being a few tokens in a fluent description. Acceptance compounds this by admitting a candidate in full, so every field of an accepted caption is imitated with equal weight and nothing pushes probability mass away from a value the model got wrong. Correcting a field calls for an objective defined on that field, scoring the value the model chose against the values it did not. A field can be isolated by prompting; what prompting cannot supply is a label trustworthy enough to optimize against. We therefore reformulate field-level verification as a question-answering problem with fixed answers, so the exam turns open-ended field correctness into verifiable supervision.

Exam construction. The exam is built from 200 human-annotated reference captions and held fixed. Each reference is probed along many of its fields rather than used once as an imitation target, yielding 3,000 four-option items that span a four-level difficulty ladder over the timbre, style, and emotion fields (Appendix D); human supervision therefore builds the verifier once instead of supervising every iteration.

Table 1: Caption quality on a human-annotated test set of 325 clips, each dimension judged against the human annotation by gemini-3-flash-preview. Columns are the 12 APS dimensions, abbreviated to their first three letters except accent (Acn) and speaking rate (Rat); AVG is their macro mean. Best in **bold**, second underlined, per column, ties share a rank. *iter0* is the zero-shot model ours is built on; MCQ-scored GSPO enters only in the last row.

Accuracy (%) ↑	**Timbre**						**Emotion**	**Style**					
Model	Gen	Age	Pit	Cla	Acn	Tim	Emo	Rat	Vol	Flu	Int	Cha	AVG
MiniCPM-o 4.5	97.7	71.4	39.5	95.5	79.3	39.6	33.7	58.1	41.2	74.3	33.6	41.3	58.8
Audio Flamingo 3	96.0	59.9	58.2	95.0	73.2	41.6	45.3	48.6	61.8	73.3	27.7	56.0	61.4
Qwen3-Omni-30B-A3B-Instruct (iter0)	98.8	68.9	55.7	97.0	90.3	47.3	57.7	66.7	73.6	75.7	56.5	46.6	69.6
MOSS-Audio-8B-Instruct	98.8	71.4	65.1	96.0	93.1	58.4	70.6	70.4	82.4	80.2	52.7	64.8	75.3
IterCap (iter1)	98.8	70.5	67.4	**97.7**	94.4	64.9	80.2	80.1	78.9	90.1	65.9	76.8	80.5
IterCap (iter2)	98.8	73.5	70.6	97.6	97.2	70.5	80.0	79.4	83.1	86.7	66.5	74.0	81.5
IterCap (iter3)	98.8	72.3	70.9	97.3	97.2	69.5	78.5	78.8	83.1	88.1	64.2	75.0	81.1
IterCap (iter4)	99.1	72.6	71.1	97.6	97.2	**72.6**	80.7	79.1	**85.9**	91.0	**69.6**	77.7	82.8
IterCap (iter5)	99.1	72.9	71.1	97.6	**98.6**	70.6	80.4	79.4	84.5	**93.3**	69.0	**78.6**	82.9
IterCap (iter5+GSPO)	**99.4**	**73.8**	**73.6**	96.3	97.2	72.4	**82.2**	**84.1**	85.7	87.8	69.3	78.4	**83.4**

Reward. Here the model is queried as an examinee: given an exam item q consisting of audio, a question, and four options, it reasons over the audio and then states its choice, so its response $\hat{o}$ terminates in one option letter, or in several when the item is multi-select. The per-item reward is exact-match correctness of that terminal choice against the reference-fixed key, $r(q, \hat{o}) = \mathbf{1}[\text{choice}(\hat{o}) = o^\star(q)]$, and the objective is the expected accuracy on the fixed exam:

$$R(\theta) \;=\; \mathbb{E}_{q\sim\mathcal{Q},\; \hat{o}\sim\pi_\theta(\cdot|q)}\big[\, r(q,\hat{o})\,\big], \tag{3}$$

where $\mathcal{Q}$ is the fixed exam and $o^\star(q)$ the reference-anchored correct option of item q. Because every answer is fixed by the reference annotation rather than by a model's assessment of a description, the reward is inexpensive and needs no judge at rollout time, unlike a learned reward model whose own errors would compound with the policy's.

Optimization. We optimize R with GSPO (Zheng et al., 2025), maximizing a clipped surrogate objective whose importance-sampling ratio is taken at the sequence level:

$$\mathcal{J}(\theta) \;=\; \mathbb{E}_{q\sim\mathcal{Q},\; \hat{o}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Big[\min\big(\rho(q,\hat{o})\,\hat{A}(q,\hat{o}),\; \text{clip}\big(\rho(q,\hat{o}),\,1{-}\epsilon_{\text{low}},\,1{+}\epsilon_{\text{high}}\big)\,\hat{A}(q,\hat{o})\big)\Big], \tag{4}$$

where $\rho(q,\hat{o}) = \big(\pi_\theta(\hat{o}\mid q)/\pi_{\theta_{\text{old}}}(\hat{o}\mid q)\big)^{1/|\hat{o}|}$ is a single length-normalized ratio over the whole response rather than a product of per-token ratios, and $\hat{A}(q,\hat{o})$ is the group-standardized advantage: for each item q, a group of G responses is sampled from the old policy, their rewards $r(q,\hat{o})$ are standardized within the group, and each response receives its group-normalized reward as its advantage. We use $G = 8$; the remaining hyperparameters are listed in Appendix G. The sequence-level ratio is what makes the objective usable on our backbone: the caption model is a mixture-of-experts model, so a token's expert routing changes between $\pi_{\theta_{\text{old}}}$ and π_θ and per-token importance weights computed across differing routing paths are unstable, whereas one ratio over the whole response is invariant to the routing.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and evaluation. We train Qwen3-TTS on the filtered subset of the ISCSLP 2026 CoT-TTS corpus (Xue et al., 2026) (Appendix C), re-annotated with IterCap captions. Instruction-following is evaluated on InstructTTSEval (Huang et al., 2025) with the official prompt at temperature 0, over 3,000 prompts per language, 1,000 each for the acoustic-parameter specification (APS), descriptive-style directive (DSD), and role-play (RP) tiers. Responses are judged by gemini-3-flash-preview, and every comparison is paired, since all systems answer the same prompts.

Baselines. We compare against five public instruct-TTS systems, namely Qwen3-TTS (Hu et al., 2026), VoxCPM2 (Zhou et al., 2026b), FireRedAudio (Li et al., 2026b), MiMo-Audio (Xiaomi,

Table 2: Caption-source ablation: instruction-following accuracy (%) on InstructTTSEval. The same Qwen3-TTS is trained on identical audio and text over the same 300K-sample corpus, so every column is a single-variable comparison. The first row is the captions the corpus ships with and the last is IterCap re-annotation of that corpus. Avg is the macro average over the six tiers; best in **bold**, second best underlined.

Accuracy (%) ↑	**InstructTTSEval-ZH**			**InstructTTSEval-EN**			
Caption source	APS	DSD	RP	APS	DSD	RP	Avg
ISCSLP (native captions)	92.8	92.9	71.0	92.9	88.8	68.5	84.5
MiniCPM-o 4.5	91.6	89.9	70.4	91.3	88.9	66.4	83.1
Audio Flamingo 3	90.7	86.3	67.0	88.9	85.7	61.8	80.1
MOSS-Audio-8B-Instruct	92.0	91.0	68.1	90.9	86.4	62.0	81.7
Qwen3-Omni-30B-A3B-Instruct	91.4	90.3	67.0	92.4	90.0	66.2	82.9
IterCap (ours)	**95.4**	**95.5**	**74.2**	**96.1**	**93.5**	**74.9**	**88.3**

2025), and Ming-omni-tts (Inclusion AI, 2026), evaluated under the same protocol. They are the most recent and strongest open-source systems that accept a free-form description at inference time. Other systems with natural-language voice control (Huang et al., 2026; Liao et al., 2026; Zhou et al., 2026a) neither report under this benchmark nor expose a caption pipeline, so we treat them as related work. An anchor row calibrates the judge's resolvable range: real recordings scored against their own human-written descriptions.

Implementation. The caption model is built on Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), trained by the flywheel of Section 3.2 and post-trained with MCQ-scored GSPO (Section 3.4). Three sets are kept disjoint: the 200 exam references, whose answers are fixed by human annotation and which are drawn from a corpus separate from the downstream one; the 100-clip development set that the round-wise check reads; and the 325-clip test set on which caption quality is scored.

4.2 Caption model quality on a human-annotated test set

Protocol. Our first evaluation measures caption correctness directly, at per-field resolution, on an internal human-annotated set of 325 clips. Each clip is annotated along the 12 attribute dimensions of InstructTTSEval's APS tier (Huang et al., 2025): gender, age, pitch, clarity, accent, timbre, emotion, speaking rate, volume, fluency, intonation, and character (Appendix C). Given the original audio, each caption model is queried separately for every dimension, and gemini-3-flash-preview judges whether its field prediction agrees with the corresponding human annotation. We report accuracy per dimension and the macro average over the 12.

Compared caption models. We compare IterCap against four open-source caption models: Audio Flamingo 3 (Ghosh et al., 2026a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), MOSS-Audio-8B-Instruct (Yang et al., 2026a), and MiniCPM-o 4.5 (Cui et al., 2026). The Qwen3-Omni entry shares IterCap's backbone and is used zero-shot, making it iteration 0 of our own model; since the compared captioners differ in scale, it is the progression from that row that holds the base model fixed. MOSS-Audio-8B-Instruct and Audio Flamingo 3 are both compared rows and the two models whose own annotations gate acceptance; our final model exceeds both on all 12 dimensions, not only on the macro average.

Results. Table 1 reports the outcome. The loop raises the macro average from 69.6 at iter0 to 82.9 at iter5, and the MCQ-scored GSPO stage adds a further 0.5 to 83.4, leaving the final model 13.8 points above the zero-shot backbone it starts from and 8.1 above MOSS-Audio. The gain concentrates where the zero-shot model is weakest, character (46.6 to 78.4), timbre (47.3 to 72.4) and emotion (57.7 to 82.2) each moving by more than 24 points, while gender and clarity, both already above 96, move by less than a point. Successive rounds contribute $+10.9$, $+1.0$, -0.4, $+1.7$ and $+0.1$, the bulk of the gain arriving in the first. What each round trained on follows the development-set check: round 1, having no earlier score to compare against, trains on its accepted annotations together with a constructed batch balancing coverage across the 12 dimensions; rounds 2 and 4 cleared the check and trained on accepted annotations alone; rounds 3 and 5 did not, and were

Table 3: Instruction-following accuracy (%) on InstructTTSEval. The human-annotation row scores real recordings paired with their human-written descriptions and marks the judge's resolvable range, not an upper bound. Best in **bold**, second underlined; Avg is the macro average over the six tiers.

Accuracy (%) ↑	**InstructTTSEval-ZH**			**InstructTTSEval-EN**			
Model	APS	DSD	RP	APS	DSD	RP	Avg
Human annotation	96.4	94.9	72.1	96.2	95.4	74.2	88.2
MiMo-Audio-7B-Instruct	86.4	85.7	68.1	90.3	86.6	66.4	80.6
VoxCPM2	94.6	93.1	67.0	94.5	91.5	70.7	85.2
Qwen3-TTS-12Hz-1.7B-VoiceDesign	95.2	**96.0**	72.2	94.1	93.8	71.6	87.2
Ming-omni-tts	**95.4**	90.4	69.9	90.1	83.6	60.1	81.6
FireRedAudio	95.3	95.2	**74.3**	95.9	**94.2**	74.7	**88.3**
Qwen3-TTS + IterCap caption (ours)	**95.4**	95.5	74.2	**96.1**	93.5	**74.9**	**88.3**

retrained with constructed audio added. Round 3 is the only round that declines on this test set, 0.4 below round 2. The first round's $+10.9$ is a combined effect, not attributable to either operation alone.

Effect of the MCQ-scored GSPO stage. Adding the exam-scored stage on top of iter5 moves the macro average from 82.9 to 83.4 and makes it the best row on five of the 12 dimensions, its largest gains on speaking rate (79.4 to 84.1) and pitch (71.1 to 73.6); four dimensions fall, fluency by 5.5 points. The stage therefore redistributes accuracy toward the fields the exam probes densely rather than improving the caption uniformly.

4.3 Caption-source ablation

A single-variable comparison. The framework's claim is about the system trained from the supervision. Architecture, audio-text pairs, data scale and training protocol are held fixed and only the caption source varies, so a difference between rows is attributable to the captions, not the model consuming them. Table 2 compares five alternative sources, among them the native chain-of-thought captions shipped with the ISCSLP corpus (Xue et al., 2026), against IterCap re-annotation of the same corpus. Replacing the native captions raises the macro average from 84.5 to 88.3, a gain of 3.8 points, and IterCap leads every column.

Effect of the backbone. The same table rules out a stronger backbone as the explanation: Qwen3-Omni-30B-A3B-Instruct, the zero-shot version of the model IterCap is built on, is itself a caption source and reaches 82.9, so the 5.4-point difference is produced by the self-annotating supervision, not the base model.

4.4 Comparison with public instruct-TTS systems

Table 3 places that controlled gain in context, comparing the same downstream system against the five public instruct-TTS systems. Our system reaches 88.3 on the macro average, on a par with the strongest public systems (FireRedAudio 88.3, Qwen3-TTS 87.2) and with the human-annotation row (88.2), which anchors the scale; it is best or tied for best on three of the six tiers. These systems differ from ours in backbone, training data and scale, so what this table adds is absolute standing, not a second causal claim; the attribution to the captions rests on the ablation above.

5 Conclusion

We formulated instruct-TTS captioning as a self-annotating problem and closed the loop with three operations: selection verifies the model's own captions and retrains on the survivors, expansion supplies delivery conditions the corpus lacks, and MCQ-scored GSPO turns field-level correctness into a verifiable reward. Together they raise caption accuracy from 69.6 to 83.4 and, with corpus and backbone fixed, downstream instruction following from 84.5 to 88.3. Appendix I discusses two limitations.

AI USE STATEMENT

Generative AI tools were used in this work in four capacities. First, as instruments of the method and evaluation: audio-language models supply the independent annotations, and a text-only model the agreement scores, for the cross-model consistency check in the flywheel (Section 3.2), and gemini-3-flash-preview is the judge used for InstructTTSEval (Section 4.1); all roles are specified in the corresponding sections. Second, for generating synthetic data: the expansion operation synthesizes training audio with Step-Audio-EditX and CosyVoice 3 (Section 3.3), and every caption the model trains on in the loop is model-generated, the exam keys being the one exception, which is the subject of the paper rather than an incidental use. Third, for implementing methods: we used AI coding assistants to write and revise parts of the training, inference and evaluation code, which we reviewed and tested. Fourth, as writing aids, to polish prose and check grammar. We did not use generative AI to generate research ideas, experimental results, or numerical values; all quantitative claims derive from experiments we ran and verified. We take full responsibility for all content of this paper, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The caption model is trained on diverse audio datasets through the self-annotating supervision flywheel (Section 3.2), with rare emotion/style injection from the controlled editing and synthesis systems named in Section 3.3 and MCQ-scored GSPO post-training (Section 3.4). The 200 human-annotated reference captions used for MCQ construction are described in Section 3.4. The resulting model re-annotates the publicly released filtered subset of the ISCSLP 2026 CoT-TTS corpus (Xue et al., 2026) (300K samples, Appendix C), and Qwen3-TTS is trained on this re-annotated corpus for instruction-following synthesis and evaluated. Appendix C specifies the typed schema and the screening step; Section 3.2 gives the acceptance predicate of Eq. (1), the instantiation of each stage, the round-wise exclusion rule, and the development-set check that decides whether a round is retrained with constructed audio, with the thresholds of both verification stages in Appendix F; Section 3.3 names the editing and synthesis systems used; and Section 3.4 defines the MCQ construction, the reward of Eq. (3), and the GSPO objective. Section 4.1 states the full evaluation protocol, including the InstructTTSEval configuration and the judge, with every system comparison paired over identical prompts. Appendix B tabulates the provenance of every surveyed caption source; Appendix D details the MCQ exam construction, Appendix E the evaluation protocol, and Appendix G the optimizer, batch, and rollout settings of both training stages. The re-annotated training corpus and the evaluation sets used in this paper will be released.

REFERENCES

Aryan Vijay Bhosale, Harshit Rajgarhia, Abhishek Mukherji, and Dinesh Manocha. Torus: A test of rendering-understanding self-coherence for unified audio models. *arXiv preprint arXiv:2607.28896*, 2026.

Jianghan Chao, Wenhui Tan, Yuchong Sun, Ruihua Song, Liyun Ru, et al. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 84035–84062, 2026.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Ensiong Chng. Audio large language models can be descriptive speech quality evaluators. In *International Conference on Learning Representations*, volume 2025, pp. 24920–24934, 2025.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *International Conference on Learning Representations*, volume 2026, pp. 24338–24362, 2026.

Cheng Siong Chin, Jianhua Zhang, and Mohan Venkateshkumar. What the waveform knows: Transparent-first speech and audio intelligence with caption studio. *arXiv preprint arXiv:2607.18704*, 2026.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. Adiff: Explaining audio difference using natural language. In *International Conference on Learning Representations*, volume 2025, pp. 11332–11365, 2025.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, et al. Compa: Addressing the gap in compositional reasoning in audio-language models. In *International Conference on Learning Representations*, volume 2024, pp. 15486–15511, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026a.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, et al. Music flamingo: Scaling music understanding in audio language models. In *International Conference on Learning Representations*, volume 2026, pp. 78958–78996, 2026b.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 4247–4271, 2025.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, volume 2026, pp. 110411–110432, 2026.

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*, 2026.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. InstructTTSEval: 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.

Inclusion AI. Ming-omni-tts: Simple and efficient unified generation of speech, music, and sound with precise control. `https://xqacmer.github.io/Ming-omni-tts/`, 2026.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J Bryan, Zeyu Jin, et al. Tac: Timestamped audio captioning. *arXiv preprint arXiv:2602.15766*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026a.

Junjie Li, Xuelong Geng, Kun Xie, Feiyu Shen, Yichen Wu, Ziqi Dai, Yichen Han, Yan Jia, Kai Huang, Junjie Chen, et al. Firereda udio: A general-purpose audio language model with decoupled continuous representations for understanding and generation. *arXiv preprint arXiv:2608.24168*, 2026b.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, et al. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*, 2026.

Zhiyu Lin, Jingwen Yang, Jiale Zhao, Meng Liu, Sunzhu Li, and Benyou Wang. Decoding the ear: A framework for objectifying expressiveness from human preference through efficient alignment. *arXiv preprint arXiv:2510.20513*, 2025.

Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jing-Shu Zheng, Zheqi He, JG Yao, Xi Yang, et al. Beyond multiple choice: Verifiable openqa for robust vision-language rft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18859–18869, 2026a.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, et al. Star-bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 134703–134731, 2026b.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Sihang Nie, Jinxin Ji, Xiaofen Xing, Deyi Tuo, Chengbin Jin, Jialong Mai, and Xiangmin Xu. Wordvoice: Explicit and decoupled multi-dimensional word-level control for llm-based tts. *arXiv preprint arXiv:2607.06461*, 2026.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Zhijiang Tang, Jiaxin Qi, Kaihua Tang, Yuhua Zheng, and Jianqiang Huang. A reconstruction-based framework for caption evaluation beyond reference captions. *arXiv preprint arXiv:2607.23235*, 2026.

Jinchuan Tian, Sang-gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, et al. Ualm: Unified audio language model for understanding, generation and reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 24563–24585, 2026.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *arXiv preprint arXiv:2601.15668*, 2026.

Daiqing Wu, Xuan Zhang, Dongbao Yang, Jiashu Yao, Longfei Chen, Qingsong Liu, Sicheng Zhao, Can Ma, Yangyang Kang, and Yu Zhou. Echo: Towards advanced audio comprehension via audio-interleaved reasoning. *arXiv preprint arXiv:2602.11909*, 2026a.

Liang-Yuan Wu, Sripathi Sridhar, Mark Cartwright, and Magdalena Fuentes. An evaluation framework for structured audio captions validated by controlled perturbations. *arXiv preprint arXiv:2607.21424*, 2026b.

LLM-Core Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR*, abs/2512.23808, 2025. doi: 10.48550/ARXIV.2512.23808. URL https://doi.org/10.48550/arXiv.2512.23808.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 13066–13093, 2026.

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

Shuhao Xu, Yifan Hu, Jingjing Wu, Zhihao Du, Zheng Lian, and Rui Liu. Emotranscap: Dataset and pipeline for emotion transition-aware speech captioning in discourses. *arXiv preprint arXiv:2604.26417*, 2026a.

Zitong Xu, Huiyu Duan, Shengyao Qin, Guangyu Yang, Guangji Ma, Xiongkuo Min, Ke Gu, Guangtao Zhai, and Patrick Le Callet. Itiscore: An image-to-text-to-image rating framework for the image captioning ability of mllms. *arXiv preprint arXiv:2604.03765*, 2026b.

Wei Xue, Junlan Feng, Shilei Zhang, Yue Wang, Ruosong Yang, Bei Liu, Liumeng Xue, Sitong Cheng, Jiahao Pan, Weizhen Bian, Boyi Kang, and Bin Long. ISCSLP 2026 CoT-TTS Challenge: Chain-of-thought reasoning for context-aware text-to-speech. *arXiv preprint arXiv:2606.21933*, 2026.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, et al. Step-audio-editx technical report. *arXiv preprint arXiv:2511.03601*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

Chenchen Yang, Kexin Huang, Liwei Fan, Qian Tu, Botian Jiang, Dong Zhang, Linqi Yin, Shimin Li, Zhaoye Fei, Qinyuan Cheng, et al. Wesr: Scaling and evaluating word-level event-speech recognition. *arXiv preprint arXiv:2601.04508*, 2026b.

Penghui Yang, Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Yibin Wang, Yujie Zhou, Jiazi Bu, Jianze Liang, Qidong Huang, et al. Caprl++: Unified reinforcement learning with verifiable rewards for dense image and video captioning. *arXiv preprint arXiv:2606.09393*, 2026c.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Mengyu Zhang, Siyu Ding, Weichong Yin, Yu Sun, and Hua Wu. Extending rlvr to open-ended tasks via verifiable multiple-choice reformulation. *arXiv preprint arXiv:2511.02463*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026a.

Xin Zhou, Yiwen Guo, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-consistency of the internal reward models improves self-rewarding language models. *arXiv preprint arXiv:2502.08922*, 2025.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, et al. Voxcpm2 technical report. *arXiv preprint arXiv:2606.06928*, 2026b.

A Extended related work

Rejection sampling fine-tuning, or RFT, samples from a model, keeps the outputs that pass a verifier, and trains on them. It descends from self-bootstrapping loops such as STaR (Zelikman et al., 2022) and Reinforced Self-Training (Gulcehre et al., 2023; Singh et al., 2023), and the term originates with Yuan et al. (2023). IterCap inherits its central limitation, that a selective operator cannot introduce behaviors absent from the output distribution, and circumvents it in two ways: controlled editing that constructs counterfactual supervision, and a two-stage verifier, the latter motivated by evidence that a model's own reward judgments are unreliable and its internal reward heads can disagree (Zhou et al., 2025).

A separate line of work conditions controllable synthesis on attribute tags or attribute embeddings rather than on language (Huang et al., 2026; Liao et al., 2026; Zhou et al., 2026a). Such an interface asks the user to know the control vocabulary and to name a value inside it, which is not how delivery is ordinarily described, and it bounds controllability by what the vocabulary enumerates; role-play directives of the kind posed by the RP tier of InstructTTSEval (Huang et al., 2025) have no tag to select. Our schema types what a caption must describe without bounding what it may say: every field is a free-form natural-language phrase rather than a value drawn from a closed set, so the interface stays compositional while remaining machine-checkable.

Decomposing captions into typed fields has served both for evaluation (Wu et al., 2026b) and for programmatically constructed data with construction-time ground truth (Kumar et al., 2026); in IterCap the schema instead makes a caption machine-checkable, as the precondition for both the clip-level consistency check in the flywheel and the field-wise multiple-choice examination. The schema is therefore a training-time constraint on a self-annotating model rather than an evaluation protocol over professional annotations. Converting open-ended captioning into a verifiable reward via multiple-choice question (MCQ) reformulation has precedent (Xing et al., 2026; Yang et al., 2026c; Zhang et al., 2025); we operate in the speech domain and address individual fields rather than the caption as a whole, and our per-field, reference-anchored construction is designed to resist

Table 4: Caption provenance of the surveyed works, grouped by the provenance regimes their captions come from; the row below the rule is a category outside the seven recurring regimes of Section 2.

Provenance regime	Works
Human annotation	CompA, MMAU, STAR-Bench, TORUS, OmniVideoBench, StructAudioCap (Ghosh et al., 2024; Sakshi et al., 2025; Liu et al., 2026b; Bhosale et al., 2026; Li et al., 2026a; Wu et al., 2026b)
Zero-shot general-purpose LALM	Omni-Captioner, UALM, Echo, Aligned better listen better, AVHBench, AVoCaDO, JointAVBench, VoxCPM2 (Ma et al., 2026; Tian et al., 2026; Wu et al., 2026a; Guo et al., 2025; Sung-Bin et al., 2025; Chen et al., 2026; Chao et al., 2026; Zhou et al., 2026b)
Model with human or signal-level verification	MOSS-Audio, Music Flamingo, EmotionThinker, SALMONN, LTU, ADIFF, ALLD, EmoTransCap, WESR (Yang et al., 2026a; Ghosh et al., 2026b; Wang et al., 2026; Tang et al., 2024; Gong et al., 2024; Deshmukh et al., 2025; Chen et al., 2025; Xu et al., 2026a; Yang et al., 2026b)
Programmatic construction	TAC, WordVoice, Caption Studio (Kumar et al., 2026; Nie et al., 2026; Chin et al., 2026)
Reinforcement learning or preference optimization	Music Flamingo, AVoCaDO, video-SALMONN 2, CapRL, CapRL++, VMR (Ghosh et al., 2026b; Chen et al., 2026; Tang et al., 2025; Xing et al., 2026; Yang et al., 2026c; Zhang et al., 2025)
Self-bootstrapping or self-iterating re-annotation	MOSS-Audio, video-SALMONN 2, and **IterCap (ours)** (Yang et al., 2026a; Tang et al., 2025)
Undisclosed	Qwen3-TTS, FireRedAudio (Hu et al., 2026; Li et al., 2026b)
Image-domain evaluation (not captioning)	ITIScore, ReconCapEval (Xu et al., 2026b; Tang et al., 2026)

the known failure modes of answering from language priors without attending to the input (He et al., 2026) or exploiting answer cues leaked across options (Liu et al., 2026a).

GSPO (Zheng et al., 2025) and its variants have recently been applied to audio reasoning (Wang et al., 2026), where a rule-based outcome reward alone is insufficient to supervise open-ended generation. Building on the MCQ reformulation noted above, we optimize this verifiable reward with GSPO: it is anchored in fixed exam answers rather than a learned judge, so it neither drifts at rollout time nor inherits the errors of a learned reward model. The same caution extends to evaluation, where LALM judges are known to exploit protocol-level shortcuts and to disagree with reconstruction- or perturbation-based tests (Chen et al., 2025; Tang et al., 2026); we therefore report the judge-based benchmark for comparability with prior work, anchored by a human-annotation row.

The captioners we compare against in Table 1 obtain their caption ability without self-iteration: Audio Flamingo 3 (Ghosh et al., 2026a) trains on captions from a single round of GPT-4.1 prompting, while Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) and MiniCPM-o 4.5 (Cui et al., 2026) caption zero-shot with their omni models.

B Caption provenance, per paper

Table 4 assigns every surveyed work to the provenance regimes that produced its captions. The first seven rows are the regimes enumerated in Section 2; the row below the rule is image-domain evaluation, which assesses captions rather than produces them and is included as a point of comparison. A work appears under more than one regime when its captions come from several sources.

C Unified caption schema and sound-event screening

The instruct-TTS training corpus is the publicly released filtered subset of the ISCSLP 2026 CoT-TTS corpus (Xue et al., 2026) (300K samples): Qwen3-TTS is trained on this subset re-annotated with IterCap captions, while the caption model itself is trained on diverse audio datasets through the flywheel (Section 3.2). This appendix specifies the typed schema that the re-annotation follows and the screening step applied to the corpus.

Table 5: An illustrative IterCap caption for one clip, in the three forms a caption carries. The typed fields fix which dimensions must be described, each holding a free-form phrase rather than a label from a closed vocabulary; the two sentence-level fields render the same delivery as free text. Values are illustrative.

Field	Value
Typed fields: the 12 APS dimensions	
Gender	female
Age	young adult
Pitch	mid-low, settling further at the end of each phrase
Clarity	clearly articulated throughout
Accent	standard Mandarin, no regional coloring
Timbre	warm and slightly breathy
Emotion	restrained disappointment, held back rather than voiced
Speaking rate	slow, with a long pause before each clause
Volume	soft, dropping further on the last clause
Fluency	fluent, without hesitation or repair
Intonation	falling contours over a narrow range
Character	composed and reserved
One-sentence description	
A young woman speaks softly and slowly in a warm, slightly breathy voice, her falling intonation and long pauses conveying a disappointment she is keeping to herself.	
Role-play instruction	
Speak as someone who has just been told something she had half expected to hear, staying composed in front of the person telling her.	

C.1 SCHEMA

Every supervision source is unified under one typed schema, which is the interface that makes heterogeneous labels machine-checkable downstream. The controllable dimensions are the 12 attribute axes defined by InstructTTSEval's acoustic-parameter specification (APS) tier (Huang et al., 2025), grouped as timbre, emotion, and style attributes, the grouping used in Table 1:

- **Timbre group**: gender, age, pitch, clarity, accent, timbre.
- **Emotion**: emotion.
- **Style group**: speaking rate, volume, fluency, intonation, character.

The schema fixes which dimensions a caption must describe, not the values it may assign them: each field holds a free-form natural-language phrase rather than a label from a closed vocabulary, and the verification stages and the exam read these phrases as text.

In addition, segment-level structure is captured through the typed `segments[].segment_type` field, which distinguishes speech from non-speech segments; this field is what the sound-event screening step reads.

Sentence-level renderings. Each caption carries, alongside the typed fields, two sentence-level renderings of the same delivery: a single free-text sentence describing the clip as a whole, and a role-play instruction phrased as a directive to a speaker. Table 5 shows an illustrative caption in all three forms.

C.2 SOUND-EVENT SCREENING

The corpus is reduced by a single screening step before the caption model iterates, distinct from the filtering already applied to the released subset. It reads the typed `segments[].segment_type` field and drops segments labeled as sound events that carry speech. Nothing else is removed, so the corpus the caption model re-annotates differs from the released subset only in these segments.

Table 6: MCQ difficulty ladder: item counts and the judgment each level requires.

Level	Items	Judgment required
L1	1,327	identification of a single field from a small candidate set
L2	725	compositional judgment over two fields
L3	222	compositional judgment over three fields
L4	726	multi-select: which of the four asserted attributes hold
Total	3,000	over the three attribute groups and their combinations

D MCQ EXAM CONSTRUCTION

The MCQ-scored GSPO reward of Section 3.4 is computed over an exam built once from a small set of human-annotated references and held fixed, so the reward does not drift at rollout time. The exam totals 3,000 four-option items drawn from 200 reference captions; each reference supports an average of fifteen independent field probes. Items span the three attribute groups, timbre, style, and emotion, and their cross-field combinations, and are distributed across the four-level difficulty ladder of Table 6.

From fields to options. Each reference caption is a natural-language description, and a parser maps its fields onto the categories the exam uses. For a discrete attribute (pitch, speaking rate, volume, clarity, fluency, accent, timbre, intonation, character, gender, and age) the correct option is the category of the reference value and the three distractors are other categories of the same attribute; the correct option is fixed by the reference annotation, and the distractors are chosen so that the item cannot be answered from language priors alone. For emotion and style the options are verbatim: the correct option is the reference sentence itself, and each distractor is a sentence taken from a different clip, sharing no trigram with it and drawn from a different valence bucket, so the item cannot be answered by lexical overlap or by the prior that a phrasing accompanies a given emotion. An illustrative single-field item reads "Which timbre attribute best matches the audio?" with options A. Raspy, B. Bright, C. Deep, D. Breathy.

Compositional and multi-select items. A compositional item asserts two or three attributes at once, and its four options are the true combination together with combinations that flip a single field. A multi-select item gives four options that each assert one attribute and marks two of them true, so its answer is the ascending string of the two correct letters. The exam therefore contains both single-answer and multiple-answer items, and the reward of Section 3.4 matches the stated choice as a whole in either case.

Safeguards. Two properties are controlled by construction. First, options are balanced per item type rather than globally: global balancing does not constrain the position distribution within a type and leaves small types skewed, concentrating the correct option of gender items at position A. Across the exam the four positions are used $572/572/568/562$ times over the $2{,}274$ single-answer items. Second, emotion references whose source annotation is marked low-confidence are excluded, because they are largely vague neutral paraphrases whose verbatim form would make a noisy key. The exam covers the 12 attribute dimensions of InstructTTSEval's APS tier (Huang et al., 2025), and is restricted to the speech attributes a caption must control.

E EVALUATION PROTOCOL

All comparisons are paired over identical prompts. Table 7 records the configuration used throughout Section 4.1.

Table 7: InstructTTSEval configuration for the downstream instruction-following evaluation.

Item	Setting
Benchmark	InstructTTSEval
Prompt	official, at temperature 0
Volume	3,000 benchmark prompts per language
Tiers	1,000 each: APS, DSD, RP
Languages	Chinese, English
Judge	gemini-3-flash-preview, temperature 0.1, thinking level low
Anchor	human annotation (real recordings + their written descriptions)

F VERIFIER INSTANTIATION

Running the independent caption models over every annotation would make multi-model captioning the dominant cost at corpus scale, so DeEAR first screens the caption's asserted fields against cheap audio statistics. Emotion and style assertions are checked against thresholded scores supplied by DeEAR, in both directions. When a caption asserts a strongly valenced emotion or style, $a_{\text{acoustic}}(x, c)$ requires the scorer's prosody > 0.6, arousal > 0.75, and naturalness > 0.5; when it asserts a flat or neutral delivery, the pair is rejected at prosody > 0.75 and arousal > 0.85. Timbre assertions are checked against frame-level signal statistics, namely fundamental frequency, energy, and speaking rate. A caption that leaves the axis unspecified clears this stage by construction, since the stage tests assertions against the signal rather than penalizing omissions, which the cross-model stage and the field-wise exam of Section 3.4 are left to detect. The thresholds are set independently per axis, each at the empirical balance point where a stricter cut admits too few and insufficiently diverse samples while a looser one filters almost nothing; we validate the choice by human evaluation, which finds over 95% of accepted annotations performance-consistent.

The cross-model stage admits any number of independent caption models; we instantiate it with two, MOSS-Audio-8B-Instruct and Audio Flamingo 3. Each is prompted with the whole clip as a single sample and with the same typed schema given to π_{θ_t}, so that it returns an annotation of its own rather than a verdict on the candidate. Qwen2.5-7B-Instruct, a text-only model that never receives the audio, then compares the candidate against each reference annotation and returns an integer agreement score from 1 to 5, where 5 denotes the same voice characterized in different words and 1 a contradiction; $a_{\text{cross}}(x, c)$ holds when both comparisons clear the round's cut. Rounds 1 to 3 require at least 4 from both models. Rounds 4 and 5 lower the cut to 3, trading per-caption precision for coverage of the clips the earlier rounds left unaccepted; this is the only setting that varies across rounds. Constructed clips pass through both stages in the same way, with one exception: the field set by the edit carries its construction-time value rather than the model's, so that field is not what the stages adjudicate, and a clip is accepted or rejected on the remaining fields.

G TRAINING CONFIGURATION

Both stages train Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) with Megatron-SWIFT in bf16 on 2×8 GPUs, with tensor parallelism 1, pipeline parallelism 2, expert parallelism 4, sequence parallelism enabled, and padding-free batching. The vision tower and the aligner are frozen throughout; the language model, the audio encoder, and the audio adapter are trained. Each round of the flywheel and the GSPO stage runs for exactly one epoch over its data.

Supervised fine-tuning (each flywheel round). Global batch size 256, micro-batch size 8, maximum sequence length 8,192, learning rate 1×10^{-5} decayed to 1×10^{-6} with a warm-up over the first 5% of iterations. Every round restarts from the base weights θ_0 and trains on the accumulated set $\mathcal{C}_t$ of Eq. (2).

MCQ-scored GSPO. Sequence-level importance sampling, group size $G = 8$, clipping range $\epsilon_{\text{low}} = 3 \times 10^{-4}$ and $\epsilon_{\text{high}} = 4 \times 10^{-4}$, no KL penalty ($\beta = 0$). Rollouts are sampled at temperature 1.0 with top-p 0.99 and a completion budget of 2,048 tokens, served by a colocated vLLM worker with tensor parallelism 4; groups whose rewards are all equal are resampled up to three times,

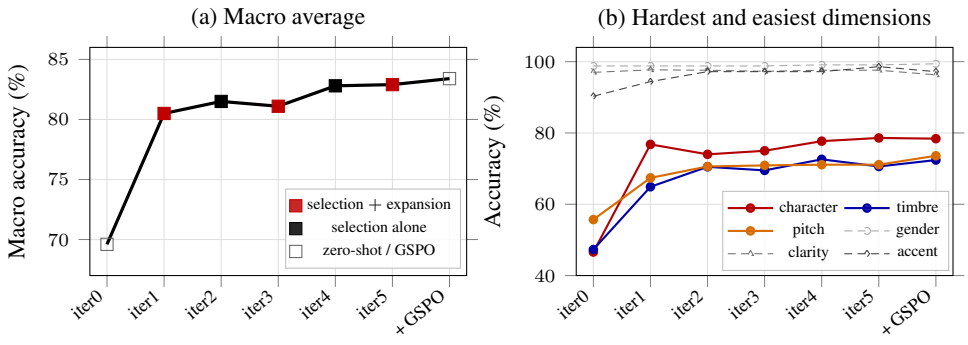

Figure 2: Per-round caption accuracy on the test set of Table 1. (a) Macro average over the 12 dimensions; marker style indicates each round's training set, rounds 1, 3 and 5 including expansion and rounds 2 and 4 trained on accepted annotations alone. Open markers denote the zero-shot initialization and the GSPO stage. (b) The three dimensions with the lowest and the three with the highest zero-shot accuracy.

and overlong rollouts are excluded from the loss. Global batch size 32 with micro-batch size 1, four optimizer steps per generation batch, learning rate 1×10^{-6} decayed to 1×10^{-7} with a warm-up over the first 3% of iterations. Alongside the exact-match reward of Eq. (3), which carries weight 1.0, a format term of weight 0.3 checks that the response states its choice in the required form; the format term is independent of which option is correct and therefore adds no answer information to the reward.

H Per-round trajectories

Figure 2 plots the rows of Table 1 against the round index, and two patterns are visible in it. The macro average rises steeply in the first round and thereafter in increments below two points, round 3 being the only decline; it is still rising at round 5, which is why we read the round count as a compute budget rather than a point of convergence. Panel (b) separates the two regimes behind that average: the three dimensions lowest at iter0 gain between 17 and 32 points over the loop, while the three highest span a range of under seven.

I Limitations

The loop's ceiling is not located. A plateau on the development set triggers remediation rather than termination, so the loop does not stop on its own. We run five rounds. The macro average in Table 1 is still rising at the last of them, but by only 0.1, so the round count reflects the compute we could allocate rather than a demonstrated upper bound, and these figures do not settle whether the loop had converged. How far further rounds would carry it, and whether the increments keep shrinking, is left to future work.

The dependence on the choice of verifier and expansion model is not measured. Selection is instantiated with MOSS-Audio-8B-Instruct and Audio Flamingo 3 as the independent caption models and Qwen2.5-7B-Instruct as the agreement judge, and expansion with the editing and synthesis systems named in Section 3.3. The framework fixes what each component must decide rather than which model decides it, but we report one instantiation: the supervision the loop can reach is bounded by what its verifiers recognize, and the capabilities expansion can introduce are bounded by what the editors can render. Quantifying either dependence requires rerunning the full five-round loop once per substituted component, which multiplies the cost of the study, so we leave it to future work. What the present results do establish is that the bound is not binding at the quality we reach: with those two verifiers the loop's captions raise downstream instruction following to 88.3, against 81.7 and 80.1 for the same system trained on those models' own captions (Table 2).