ELICITING LATENT CAPABILITIES IN SPOKEN DIALOGUE MODELS VIA ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training has become increasingly important for unlocking and refining the capabilities of foundation models. Spoken dialogue models introduce an additional challenge, as post-training should improve semantic reasoning and paralinguistic responsiveness jointly. We find that both are limited not by a lack of capability but by insufficient elicitation from speech alone. Providing transcripts and structured rationales substantially recovers both. Further analysis reveals local redundancy in discrete audio representations. Based on these findings, we propose **VoxOPD**, a cross-modal hierarchical on-policy self-distillation framework for spoken dialogue models. During training, the student generates response trajectories from speech alone, while a frozen teacher initialized from the same model conditions on privileged context to provide dense supervision along those trajectories. Beyond token-level distribution alignment, the framework assigns shared credit to audio segments through chunk-level supervision, accommodating differences in information density and local redundancy between text and audio representations. This hierarchical supervision transfers the capabilities elicited under augmented contexts to the speech-only setting. On Step-Audio 2 mini, VoxOPD improves the VoiceBench capability-specific task average by 4.0 and VStyle by 0.71. Further experiments on Kimi-Audio and VITA-Audio consistently demonstrate its effectiveness in improving both reasoning and spoken interaction.

1 INTRODUCTION

End-to-end speech dialogue models are extending human-computer interaction beyond text-based conversation to direct spoken communication. Compared with cascaded systems that combine automatic speech recognition, a language model, and speech synthesis, these models jointly capture lexical semantics and paralinguistic information such as emotion and prosody, providing a unified framework for speech-based reasoning and interaction. However, prior work has shown that existing models still face two key limitations: speech representations do not fully activate the knowledge and reasoning capabilities of the backbone language model (Chen et al., 2026b; Xiang et al., 2025; Wang et al., 2026a; Hu et al., 2026), and paralinguistic cues in the input do not stably modulate response content and spoken expression (Lin et al., 2024; Yang et al., 2026; Zhou et al., 2026).

We further verify and illustrate these limitations by controllably varying the input conditions of the same model. As shown in Figure 1(a), compared with speech-only input, both text-only input and speech paired with its transcript improve performance on semantic reasoning tasks, with further gains when a structured rationale is added. This indicates that even when the model correctly recognizes the spoken content, speech representations may still fail to activate the internal knowledge and reasoning capabilities of the language model as effectively as text tokens. We refer to this phenomenon as the *modality-conditioned capability gap*. Beyond lexical semantics, speech also carries paralinguistic signals that transcripts cannot express: on paralinguistic response tasks, conditions with speech input significantly outperform those without it, showing that paralinguistic cues provide additional information for response generation. However, these cues do not induce stable, directionally consistent changes in the model's response distribution, and therefore fail to reliably trigger corresponding adjustments in content planning and spoken expression. We refer to this phenomenon as *paralinguistic response mapping instability*. When the model is provided with an explicit rationale, response appropriateness and expressiveness improve further, indicating that the relevant capabilities are not entirely absent but require guidance to be reliably elicited.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

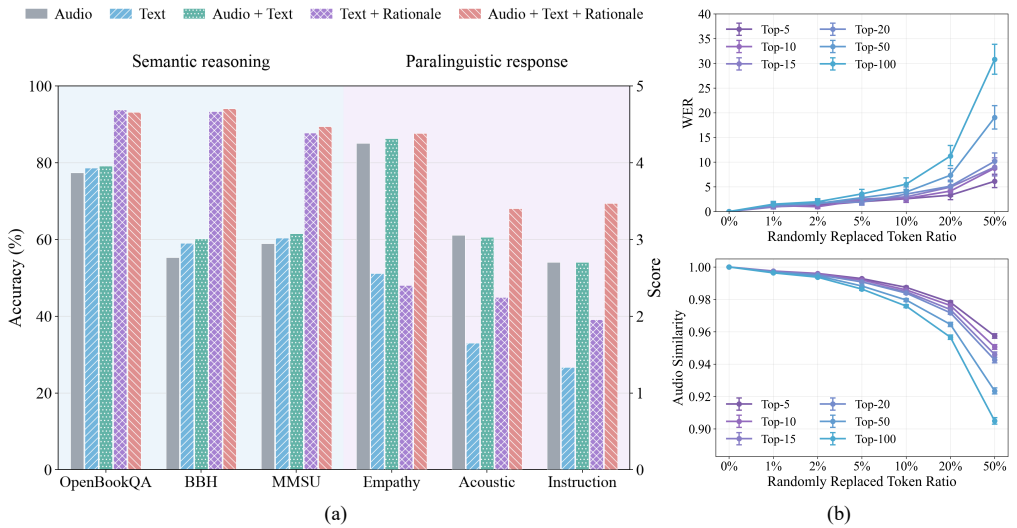


Figure 1: (a) Controlled input comparison. Each rationale contains no target answer and provides only reasoning cues to guide the model. (b) Audio-token perturbation experiment. We randomly replace audio tokens at different ratios, decode the perturbed token sequences, and compare the resulting audio with the originals. Further experimental details are provided in Appendix A.

Existing efforts to address these gaps through post-training mainly rely on supervised fine-tuning (SFT) (Wang et al., 2023; 2024) or reinforcement learning (RL) (Zhang et al., 2024; Lin et al., 2025; Kim et al., 2026). Although both paradigms have achieved respective success, they still face inherent limitations. SFT learns from offline reference responses along teacher-forced trajectories, leading to a discrepancy between training and inference (Agarwal et al., 2024; Gu et al., 2023). RL can optimize trajectories generated by the model, but its effectiveness depends critically on the quality and granularity of the reward signal. For open-ended speech responses, reliable rewards are difficult to construct, while sparse sequence-level feedback provides inadequate credit assignment over long audio-token sequences (Ji et al., 2025; Lin et al., 2026). These limitations call for dense supervision on trajectories sampled from the deployment policy itself. Taken together with the experimental evidence above, we propose that spoken dialogue models are naturally suited to on-policy self-distillation (OPSD), which transfers their latent capabilities along self-generated trajectories.

Although OPSD has proven effective across diverse domains (Zhao et al., 2026; Yuan et al., 2026; Li et al., 2026; Jiang et al., 2026; Liu et al., 2026), directly applying its token-level formulation to speech overlooks a key modality difference. Text tokens typically carry local semantic decisions directly, whereas audio tokens have higher temporal density and greater local redundancy. As shown in Figure 1(b), randomly replacing 2% of the audio tokens with samples from their respective Top-K neighbors leaves the decoded speech nearly unchanged, while increasing the rate to 5% produces only minor differences. These results suggest that local redundancy can absorb per-token distributional deviations without meaningfully altering the decoded output. Token-level alignment may therefore overemphasize output-insensitive differences and yield limited gains in actual generations. This motivates complementary chunk-level supervision that assigns shared credit to neighboring tokens and captures their collective acoustic behavior.

Motivated by these findings, we propose **VoxOPD**, a hierarchical on-policy self-distillation framework for post-training end-to-end spoken dialogue models. The student generates responses conditioned solely on user speech, while a frozen privileged teacher, initialized from the same model, additionally has access to the transcript and a structured rationale, providing dense supervision at every position along the student-generated trajectory. Privileged context is only provided to the teacher during training, ensuring that the deployed student relies on speech alone. Our hierarchical objective combines fine-grained distributional distillation over text and audio tokens with chunk-level supervision that assigns shared credit to audio segments, matching the local redundancy observed above. On our primary model, Step-Audio 2 mini, VoxOPD raises the VoiceBench capability-specific task average from 60.7 to 64.7 and VStyle from 3.14 to 3.85 relative to the original model, outperforming SFT, GRPO, and representative on-policy distillation methods trained on the same data.

Our main contributions are as follows:

- We verify the modality-conditioned capability gap and paralinguistic response mapping instability through controlled input comparisons and demonstrate local redundancy in audio representations through audio-token perturbation experiments.
- We propose VoxOPD, a cross-modal hierarchical on-policy self-distillation framework that distills semantic reasoning and paralinguistic response capabilities from augmented contexts to a spoken dialogue model operating under speech-only conditions.
- Extensive experiments across Step-Audio 2 mini, Kimi-Audio, and VITA-Audio demonstrate the effectiveness and transferability of VoxOPD, while further analyses reveal the complementary roles of privileged contexts and hierarchical supervision.

2 Related Work

Spoken Dialogue Models and Modality Gaps. Cascaded spoken dialogue systems reuse text-model capabilities but lose paralinguistic cues at the transcript interface (Huang et al., 2024; Défossez et al., 2024). End-to-end models directly take speech as input and generate spoken responses within a unified architecture (Zeng et al., 2024; Wu et al., 2025; Ding et al., 2025), yet still exhibit a modality-conditioned capability gap in knowledge and reasoning across speech and text inputs (Chen et al., 2026b; Wang et al., 2026a; Hu et al., 2026), and paralinguistic response mapping instability, failing to translate perceived emotion or prosody into appropriate content and style (Yang et al., 2026; Zhou et al., 2026; Zhan et al., 2026). Existing approaches use paired transcripts as text anchors or emotion annotations as paralinguistic supervision to narrow these gaps (Wang et al., 2026a; Kim et al., 2026), but reliably translating paralinguistic cues into appropriate response content and style remains challenging (Yang et al., 2026; Zhou et al., 2026; Zhan et al., 2026).

SFT and RL for Spoken Dialogue Models. SFT offers dense supervision for each token but suffers from teacher-forcing mismatch and collapses diverse valid responses into one target. RL instead optimizes trajectories generated by the model, typically with one reward for the entire sequence (Yang et al., 2026; Shao et al., 2024; Zhang et al., 2024; Lin et al., 2025). However, the quality of an open-ended spoken reply is hard to capture with a single scalar. One line of work improves reward reliability through generalist evaluators that jointly assess semantic quality and paralinguistic expression, or derives rewards from explainable reasoning about emotion and prosody (Ji et al., 2025; Wang et al., 2026b). Another line combines SFT and RL, using dense token-level supervision to anchor the policy and stabilize sequence-level reward optimization while learning from model-generated trajectories (Chen et al., 2026a; Lin et al., 2026), yet their differing objectives make the two supervision signals difficult to balance.

On-Policy Distillation. On-policy distillation applies token-level distributional supervision to trajectories generated by the student (Agarwal et al., 2024; Gu et al., 2023). Privileged self-distillation uses a single model as both teacher and student, with references, rationales, or private facts available only to the teacher (Zhao et al., 2026; Zhu et al., 2026; Penaloza et al., 2026; Lazaridis et al., 2026). Existing distillation methods for speech typically use text as the teacher and supervise text outputs at token or trajectory level, leaving speech realization only indirectly constrained (Wang et al., 2026c; Hu et al., 2026; Cao et al., 2026; Fu et al., 2026). We instead perform self-distillation on mixed text–audio trajectories that match deployment, giving the teacher a transcript and a structured rationale as privileged context. Beyond token-level distributional alignment, we introduce shared credit within audio chunks to leverage local acoustic redundancy.

3 Method

3.1 Problem Formulation

We define the training set as $\mathcal{D} = \{(a_i, x_i, r_i)\}_{i=1}^{N}$, where a_i is the user speech, x_i is its transcript, and r_i is query-specific guidance in the form of either a reference reasoning path or an analysis of paralinguistic cues. The transcript and guidance are available only to the teacher during training. We denote the student and the privileged teacher by π_θ and π_T, respectively. Both are initialized from

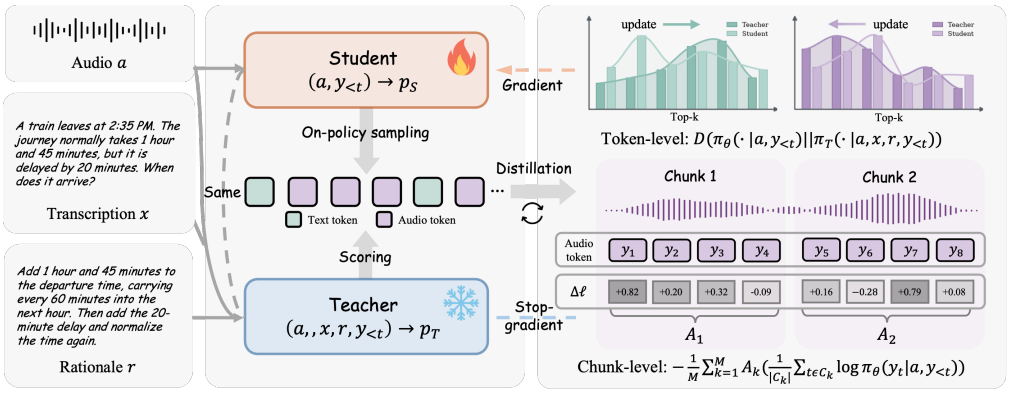


Figure 2: Overview of VoxOPD. The student samples a trajectory from speech alone and a frozen privileged teacher scores it. Token-level distillation aligns semantic and acoustic predictions, whereas chunk-level advantages assign shared credit across audio tokens.

the same pretrained model. The teacher remains frozen throughout training, and only the student parameters θ are updated. Given an example (a, x, r), the two policies condition on $\pi_\theta(\cdot \mid a)$ and $\pi_T(\cdot \mid a, x, r)$, respectively. The transcript clarifies the linguistic content, while the guidance informs the formulation of an appropriate response.

For each training example, the student first samples a response autoregressively from its current policy, conditioned only on the speech input:

$$y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid a). \tag{1}$$

In an end-to-end spoken dialogue model, y is a mixed sequence of text tokens, audio tokens, and a small number of control tokens. At every position t, both policies predict the next token from the same generated prefix $y_{<t}$. Their only difference is that the teacher additionally observes privileged information available during training. Their predictive distributions at this position are

$$\pi_\theta(v \mid a, y_{<t}), \qquad \pi_T(v \mid a, x, r, y_{<t}), \tag{2}$$

for the student and teacher, respectively. The teacher does not generate a separate reference trajectory. Instead, it provides next-token supervision at prefixes sampled from the student policy. Unlike distillation on fixed reference responses, this design trains the student on the states it actually visits and enables direct comparison between the teacher and student distributions at each shared prefix.

We partition the valid positions according to the type of generated token:

$$\mathcal{I}_{\text{text}} = \{t : y_t \text{ is a text token}\}, \qquad \mathcal{I}_{\text{audio}} = \{t : y_t \text{ is an audio token}\}. \tag{3}$$

This partition allows us to tailor supervision to each modality: text tokens are supervised individually to capture semantic decisions, whereas audio tokens are supervised at both the token and chunk levels to account for their dense, locally correlated structure.

3.2 Token-Level Cross-Modal Distillation

We first provide token-level supervision through separate text and audio losses, aligning the student and teacher predictions for individual semantic and acoustic decisions.

At both text and audio positions, we adopt the Top-K stop-gradient reverse-KL objective of Zhu et al. (2026). Restricting the vocabulary support reduces the cost of distribution matching. The stop-gradient operation removes the extra gradient term induced by unnormalized Top-K truncation. For modality m, we define

$$\mathcal{L}_m = -\frac{1}{|\mathcal{I}_m|} \sum_{t \in \mathcal{I}_m} \sum_{v \in \mathcal{S}_t^K} \pi_\theta(v \mid a, y_{<t}) \Big[\log \pi_T(v \mid a, x, r, y_{<t}) - \text{sg}(\log \pi_\theta(v \mid a, y_{<t})) \Big], \qquad m \in \{\text{text}, \text{audio}\}. \tag{4}$$

Here, $\mathcal{S}_t^K$ denotes the Top-K candidate set at position t. We construct this set from the teacher distribution by retaining its K highest-probability tokens. The student probability weighting each term remains differentiable, whereas $\mathrm{sg}(\cdot)$ treats the student log-probability in the log-ratio between the teacher and student distributions as constant. The resulting log-ratio acts as an advantage-like coefficient for updating the student. Normalizing each loss by $|\mathcal{I}_m|$ prevents the denser audio sequence from dominating merely because it contains more tokens.

The text loss $\mathcal{L}_{\text{text}}$ transfers the teacher's semantic preferences and response strategies, whereas $\mathcal{L}_{\text{audio}}$ provides fine-grained alignment of acoustic predictions throughout the sampled response.

3.3 Chunk-Level Audio Credit Assignment

In spoken responses, both semantic content and paralinguistic attributes, including prosody, rhythm, and speaking style, are realized across multiple audio tokens. Token-level supervision provides fine-grained feedback but does not explicitly share credit across neighboring tokens, whereas a single utterance-level score may be too coarse to guide local audio decisions.

We therefore partition audio tokens into non-overlapping chunks and use the average teacher–student log-likelihood gap within each chunk as a shared credit signal. This objective complements $\mathcal{L}_{\text{audio}}$. The token-level loss maintains fine-grained distributional alignment, while the chunk-level loss couples updates across neighboring audio tokens.

Let $T_a = |\mathcal{I}_{\text{audio}}|$ denote the number of audio tokens, and let $(t_1, \ldots, t_{T_a})$ list their positions in generation order. Given a fixed chunk size c, we form $M = \lceil T_a/c \rceil$ non-overlapping chunks $\mathcal{C} = \{C_k\}_{k=1}^{M}$. Specifically, $C_k = (t_{k,1}, \ldots, t_{k,m_k})$, where $t_{k,j} = t_{(k-1)c+j}$ and $m_k = \min\{c, T_a - (k-1)c\}$. Thus, each chunk contains c audio tokens except possibly the final chunk. We first compute the difference between the teacher and student log-likelihoods for each sampled token and then average within the chunk:

$$A_k = \mathrm{sg}\left[\frac{1}{m_k}\sum_{j=1}^{m_k}\left(\log \pi_T\big(y_{t_{k,j}} \mid a, x, r, y_{<t_{k,j}}\big) - \log \pi_\theta\big(y_{t_{k,j}} \mid a, y_{<t_{k,j}}\big)\right)\right]. \tag{5}$$

Thus, A_k is the stop-gradient mean log-likelihood ratio of the sampled chunk: it is positive when the privileged teacher assigns the sampled audio realization higher likelihood than the student. Using this shared signal, we define the policy-gradient surrogate

$$\mathcal{L}_{\text{chunk}} = -\frac{1}{M}\sum_{k=1}^{M} A_k \left[\frac{1}{m_k}\sum_{j=1}^{m_k}\log \pi_\theta\big(y_{t_{k,j}} \mid a, y_{<t_{k,j}}\big)\right]. \tag{6}$$

Because gradients are stopped through A_k, they propagate only through the student log-probabilities in Eq. 6. Consequently, every token in a chunk receives the same credit. A positive A_k increases the chunk's average log-likelihood, whereas a negative value decreases it. Averaging over tokens makes the scale comparable across chunks of different lengths. The chunk-level objective couples neighboring audio decisions while retaining the fine-grained alignment supplied by $\mathcal{L}_{\text{audio}}$.

The final training objective is

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{audio}}\mathcal{L}_{\text{audio}} + \lambda_{\text{chunk}}\mathcal{L}_{\text{chunk}}, \tag{7}$$

where λ_{text}, λ_{audio}, and λ_{chunk} control the relative strengths of the three supervision granularities. We set $\lambda_{\text{text}} = \lambda_{\text{audio}} = 1$ by default and treat λ_{chunk} as a tunable hyperparameter.

4 Experiments

4.1 Experimental Setup

Models. We conduct experiments on three publicly available models: Step-Audio 2 mini (Wu et al., 2025), Kimi-Audio (Ding et al., 2025), and VITA-Audio (Long et al., 2025). We use Step-Audio 2 mini as our primary model and tune the training configuration exclusively on it. We then apply the same configuration to Kimi-Audio and VITA-Audio without model-specific hyperparameter tuning, allowing us to assess the method's transferability across architectures. Detailed training settings are provided in Appendix B.

Table 1: Results on VoiceBench. The two Avg. columns report the average performance on open-ended dialogue and capability-specific tasks, respectively. Higher is better.

Model / Method	Alpaca	Common	Wild	Avg.	SD-QA	MMSU	OBQA	BBH	IFEval	Adv.	Avg.
Interleaved Text–Audio Token Generation											
Step-Audio 2	2.87	3.34	2.72	2.98	42.5	59.0	77.4	55.3	44.7	85.4	60.7
+ SFT	3.57	**3.75**	3.04	**3.45**	43.2	58.4	81.5	59.3	39.3	93.1	62.5
+ GRPO	2.11	3.10	2.04	2.41	43.3	60.6	81.1	59.7	36.9	83.1	60.8
+ CORD	3.20	3.29	2.75	3.08	42.2	59.7	77.8	59.4	**44.9**	82.7	61.1
+ ParaBridge	2.45	2.62	1.86	2.31	39.5	54.8	75.8	57.5	39.8	90.8	59.7
+ VoxOPD	**3.59**	3.63	**3.05**	3.42	**43.5**	**61.4**	**82.6**	**60.5**	**44.9**	**95.0**	**64.7**
Hybrid Discrete–Continuous Audio Representation											
Kimi-Audio	3.68	3.65	3.29	3.54	60.7	60.7	83.5	67.8	61.2	**100.0**	72.3
+ SFT	**3.85**	**3.95**	3.31	**3.70**	52.7	53.0	78.2	57.8	**64.2**	**100.0**	67.6
+ GRPO	3.70	3.61	3.38	3.56	61.0	**61.2**	83.7	67.1	61.6	**100.0**	72.4
+ VoxOPD	3.81	3.76	**3.41**	3.66	**61.6**	60.2	**85.5**	**68.8**	58.9	**100.0**	**72.5**
Multiple Cross-modal Token Prediction											
VITA-Audio	3.90	3.61	3.06	3.52	39.2	46.5	76.0	53.2	38.5	98.3	58.6
+ SFT	3.86	3.59	3.05	3.50	38.8	44.0	74.9	52.6	**49.0**	99.2	59.8
+ GRPO	**3.91**	3.58	3.01	3.50	39.2	46.2	75.4	52.6	36.3	99.0	58.1
+ VoxOPD	**3.91**	**3.66**	**3.13**	**3.57**	**39.5**	**46.9**	**76.5**	**54.3**	47.9	**100.0**	**60.9**

Benchmarks. We evaluate on VoiceBench (Chen et al., 2026b), which assesses content understanding and semantic responsiveness, and VStyle (Zhan et al., 2026), which assesses the use of paralinguistic information and the expressiveness of generated speech. The open-ended dialogue average (VB-Open) covers AlpacaEval, CommonEval, and WildVoice, while the capability-specific task average (VB-Cap) covers SD-QA, OpenBookQA, MMSU, BBH, IFEval, and AdvBench. Detailed definitions are provided in Appendix C.

Baselines. For the comparison of post-training methods, we adopt supervised fine-tuning (SFT) and group relative policy optimization (GRPO) (Shao et al., 2024) as baselines on all three models. On Step-Audio 2 mini, we further include CORD (Hu et al., 2026) and ParaBridge (Wang et al., 2026c) as representative on-policy distillation methods targeting the modality reasoning gap and paralinguistic response, respectively. All methods use the same training-data pool. Details of the data construction process are provided in Appendix E.

4.2 Main Results

Improved Semantic Understanding and Reasoning. As shown in Table 1, our method raises the open-ended dialogue average for Step-Audio 2 mini from 2.98 to 3.42 and the overall capability-specific task score from 60.7% to 64.7%. SFT achieves a comparable open-ended dialogue average of 3.45 but a lower overall capability-specific task score of 62.5%, suggesting that teacher forcing on offline references does not fully correct the states encountered during free generation. GRPO improves knowledge and reasoning scores, but leaves the overall capability-specific task score nearly unchanged and degrades open-ended dialogue, reflecting the limited and sparse supervision provided by a sequence-level scalar reward. The specialized on-policy distillation baselines are also less balanced: CORD yields only modest gains on both aggregate metrics, whereas ParaBridge performs below the base model on both. CORD conditions its teacher on pure text and uses the resulting output as both the distillation target and reward reference, leaving errors in the text-conditioned teacher uncorrected. ParaBridge distills the effect of a generic paralinguistic scaffold, providing no direct semantic or reasoning enhancement. Our frozen teacher instead retains the speech input and receives its transcript and query-specific guidance, yielding a more informative distributional target.

Stronger Paralinguistic Expressiveness. Table 2 shows that our method raises the Step-Audio 2 mini VStyle average from 3.14 to 3.85. These consistent gains suggest that the model learns a robust paralinguistic response mapping rather than overfitting to a single type of expressive instruction. In contrast, GRPO scores 3.10, slightly below the base-model score of 3.14. We attribute this result to the limited ability of a scalar reward to represent multidimensional paralinguistic quality. SFT

raises the average to 3.74 but trails our method on every dimension except role play. This gap may stem from single-reference supervision, which underrepresents the range of acceptable acoustic realizations for the same content. CORD and ParaBridge yield only marginal gains, reaching 3.17 and 3.20, respectively. The text-only teacher in CORD lacks access to paralinguistic cues, whereas the generic scaffold in ParaBridge highlights such cues without explicitly guiding how they should shape the generated speech. Both methods also lack dedicated supervision on the audio modality. Our method instead applies token- and chunk-level supervision directly to audio tokens, capturing expressive properties that unfold over short acoustic spans.

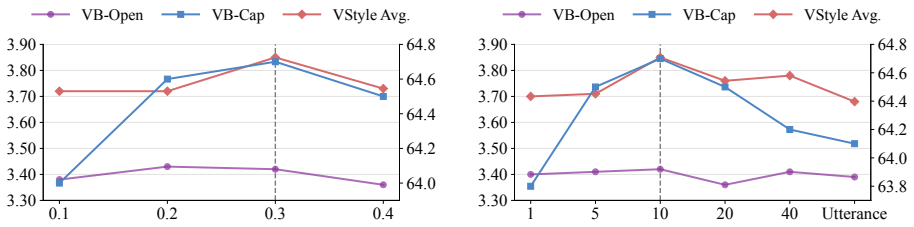
Table 2: Results on VStyle. The four dimensions are acoustic performance (Acous.), style control (Instr.), role play (Role), and empathetic response (Emp.), each scored on a 1–5 scale. Avg. reports their mean. Higher is better.

Model / Method	Acous.	Instr.	Role	Emp.	Avg.
Interleaved Text–Audio Token Generation					
Step-Audio 2	3.06	2.70	2.55	4.25	3.14
+ SFT	3.52	3.43	**3.52**	4.50	3.74
+ GRPO	2.78	2.95	2.30	4.38	3.10
+ CORD	3.05	2.82	2.45	4.38	3.17
+ ParaBridge	3.25	3.50	2.14	3.92	3.20
+ VoxOPD	**3.64**	**3.65**	3.42	**4.69**	**3.85**
Hybrid Discrete–Continuous Audio Representation					
Kimi-Audio	3.44	3.23	3.18	4.22	3.51
+ SFT	3.15	3.15	**3.80**	**4.68**	3.70
+ GRPO	**3.84**	3.65	3.48	4.33	**3.82**
+ VoxOPD	3.65	**3.79**	3.51	4.31	**3.82**
Multiple Cross-modal Token Prediction					
VITA-Audio	2.59	2.96	2.46	4.72	3.18
+ SFT	2.89	2.80	**2.92**	4.79	**3.35**
+ GRPO	2.78	**3.05**	2.41	4.70	3.23
+ VoxOPD	**2.93**	2.57	**2.92**	**4.85**	3.32

Cross-Architecture Transferability. Post-training on Kimi-Audio and VITA-Audio reuses the configuration selected for Step-Audio 2 mini, without model-specific hyperparameter search. The three models adopt interleaved text–audio token generation, hybrid discrete–continuous audio representations, and multiple cross-modal token prediction, respectively. Despite these differences, applying our method to both Kimi-Audio and VITA-Audio yields improvements in several aggregate metrics, including open-ended dialogue, capability-specific tasks, and VStyle. This broadly consistent trend across the two backbones suggests that the benefits of our method are not confined to a particular model or evaluation dimension. SFT slightly outperforms our method on Kimi-Audio open-ended dialogue and on VITA-Audio VStyle, but lowers the Kimi-Audio capability-specific task score from 72.3% to 67.6% and reduces the VITA-Audio open-ended dialogue average below the base model. On Kimi-Audio, GRPO attains the same VStyle average of 3.82, but yields only limited gains on open-ended dialogue and capability-specific tasks. On VITA-Audio, GRPO falls below our method on all three aggregate metrics. These results indicate that the gains do not rely on architecture-specific tuning and cover both semantic and paralinguistic abilities under different audio representations and generation paradigms.

4.3 Ablation Study

Complementary Supervision Signals. The three terms in the hierarchical objective provide complementary signals for different capabilities. As shown in Table 3, $\mathcal{L}_{\text{text}}$ alone raises the open-ended dialogue and capability-specific task averages from 2.98 and 60.7 for the base model to 3.41 and 64.2. This recovers most of the semantic gains of the full objective, suggesting that text-channel distribution alignment accounts for much of the reduction in the modality-conditioned capability gap. Yet its VStyle average remains at 3.52, compared with 3.85 for the full objective, indicating that semantic alignment alone is insufficient for strong paralinguistic expressiveness. Adding $\mathcal{L}_{\text{audio}}$ raises VStyle from 3.52 to 3.71 while leaving the semantic metrics broadly comparable at 3.39 and 63.9, indicating that direct audio-channel distribution supervision primarily improves expressiveness. Adding $\mathcal{L}_{\text{chunk}}$ instead yields 3.41, 64.3, and 3.70 on the three aggregates, providing VStyle gains comparable to those from $\mathcal{L}_{\text{audio}}$ while maintaining the semantic improvements of text supervision. Combining both forms of audio-side supervision achieves the strongest results, with scores of 3.42, 64.7, and 3.85. The rise in the capability-specific task score from 64.2 under text-only supervision to 64.7 under the full objective further suggests that understanding and reasoning can benefit from audio-side supervision when it is combined with text alignment. Overall, these results show that direct audio distribution alignment and chunk-level supervision provide complementary rather than redundant signals across semantic and paralinguistic capabilities.



(a) Chunk-loss weight λ_{chunk}.

(b) Audio chunk size c.

Figure 3: Sensitivity to hyperparameters on VoiceBench and VStyle. In both panels, the left axis reports VB-Open and the VStyle average, while the right axis reports VB-Cap. Dashed lines indicate the default settings. (a) Effect of the chunk-loss weight, with $\lambda_{\text{chunk}} = 0.3$ by default. (b) Effect of the audio chunk size, with a default of 10 audio tokens.

Hyperparameter Sensitivity. Figure 3(a) shows that VB-Open changes only modestly across the tested values of λ_{chunk}, while VB-Cap and VStyle achieve their highest observed scores at the default setting of $\lambda_{\text{chunk}} = 0.3$. Figure 3(b) examines the effect of credit-sharing granularity by comparing single-token credit assignment, chunks of different lengths, and utterance-level aggregation. When $c = 1$, the auxiliary objective remains active, but no credit is shared across tokens; under utterance-level aggregation, all audio tokens share a single aggregate signal. Among the settings considered, $c = 10$ achieves the highest VB-Cap and VStyle scores, outperforming both endpoint settings, while VB-Open varies only slightly. These results indicate that, under the current objective and training configuration, an intermediate granularity of local credit sharing performs best empirically, whereas extending the supervision scope to the entire utterance provides no further benefit. This trend is consistent with the design motivation of sharing feedback across neighboring decisions while avoiding signal mixing over excessively broad spans.

Table 3: Ablation study of the hierarchical objective components on Step-Audio 2 mini.

Objective	VB-Open	VB-Cap	VStyle
$\mathcal{L}_{\text{text}}$	3.41	64.2	3.52
$\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}}$	3.39	63.9	3.71
$\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{chunk}}$	3.41	64.3	3.70
$\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{chunk}}$	**3.42**	**64.7**	**3.85**

4.4 Analysis

Privileged Information. Figure 4 disentangles the roles of different forms of teacher-side privileged information. Relative to the base model, distillation from a teacher whose privileged input is limited to the transcript improves all three semantic reasoning tasks and yields modest gains on the three paralinguistic response metrics. These results show that the transcript alone can serve as effective teacher-side privileged information by eliciting capabilities that speech-only conditioning activates inconsistently and making them available for transfer through distillation. Providing the teacher with the original speech in addition to the transcript yields further gains because speech retains paralinguistic cues absent from text. Due to this setting requires no rationale construction, it offers a practical alternative when annotation or generation budgets are limited. Adding a query-specific rationale to the speech input improves the teacher's overall performance, and distillation transfers part of this gain to the speech-only student. Nonetheless, the resulting student consistently underperforms VoxOPD across all six metrics. Since the VoxOPD teacher additionally receives the transcript, this gap indicates that a rationale cannot fully substitute for the complementary lexical evidence provided by an explicit textual representation of the spoken content, which helps elicit capabilities not reliably activated by speech and rationale alone.

Modality-Specific Alignment. We further examine how the supervision terms affect modality-specific alignment on the states that the student actually visits. At different stages of training, we let the student generate responses to speech inputs from a held-out evaluation set disjoint from the training data and use the teacher to score the resulting trajectories. As shown in Figure 5, a higher overlap-token advantage and a lower absolute entropy gap indicate closer alignment. With $\mathcal{L}_{\text{text}}$ included in all four configurations, each exhibits a rapid increase in text-token advantage accompanied

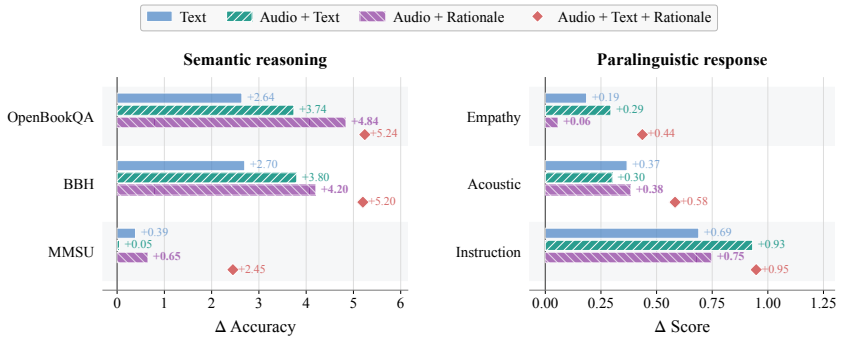


Figure 4: Performance gains over the base Step-Audio 2 mini after distillation, evaluated on students that receive speech-only input. All variants use the full training objective and differ only in the teacher input condition.

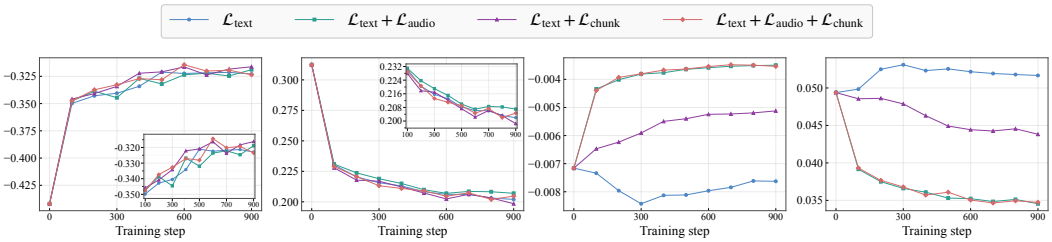


Figure 5: On-policy alignment between the teacher and student during training, evaluated on trajectories generated by the student from held-out speech inputs. Panels (a, b) report text-token alignment, and panels (c, d) report audio-token alignment. Panels (a, c) show overlap-token advantage (higher is better), whereas panels (b, d) show absolute entropy gap (lower is better).

by a reduction in the entropy gap. Adding $\mathcal{L}_{\text{chunk}}$ further improves text-channel alignment, supporting the view that meaningful semantic content in speech typically spans multiple audio tokens. The audio channel exhibits a clear contrast. With $\mathcal{L}_{\text{text}}$ alone, the audio-token advantage remains low and the entropy gap instead widens, indicating that text-only distillation has limited ability to improve output expressiveness indirectly, consistent with the results in Table 3. Adding $\mathcal{L}_{\text{audio}}$ markedly improves both audio-alignment metrics, whereas adding only $\mathcal{L}_{\text{chunk}}$ to $\mathcal{L}_{\text{text}}$ closes the alignment gap only partially. Compared with the $\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}}$ configuration, the full objective achieves a similar level of token-level audio alignment but a higher VStyle score in the ablation study. Overall, $\mathcal{L}_{\text{audio}}$ primarily aligns token-level distributions in the audio channel, whereas $\mathcal{L}_{\text{chunk}}$ complements this fine-grained signal by assigning shared credit across spans of locally redundant audio tokens. Its contribution is therefore reflected more strongly in the higher VStyle score than in further gains in token-level alignment.

5 Conclusion

We presented VoxOPD, a hierarchical cross-modal on-policy self-distillation framework for end-to-end spoken dialogue models. VoxOPD transfers capabilities elicited by privileged transcripts and structured rationales to a speech-only student. During training, a frozen teacher sharing the student's initialization scores student-generated trajectories, combining token-level text and audio distillation with chunk-level credit assignment. Across Step-Audio 2 mini, Kimi-Audio, and VITA-Audio, VoxOPD consistently improves semantic and paralinguistic performance. On Step-Audio 2 mini, it raises the VoiceBench capability-specific average from 60.7 to 64.7 and VStyle from 3.14 to 3.85, outperforming SFT, GRPO, and specialized distillation baselines overall. Ablations show that text alignment strengthens semantic capabilities, whereas audio alignment and chunk-level credit improve expressive generation. Overall, VoxOPD provides an effective approach to jointly improving reasoning and expression in spoken dialogue models.

AI USE STATEMENT

In this work, generative AI tools formed part of our core methodology for data construction. Specifically, we employed large language models and text-to-speech models, including Gemini-3-Flash, CosyVoice3, IndexTTS 2.5, CosyVoice2, Qwen3-TTS, and VoxCPM2, to filter prompts for speech suitability, categorize tasks, render visually dependent expressions in speakable form, generate reference answers and rationales, and synthesize speech. We also used ChatGPT and Claude for minor stylistic and grammatical revisions and for assistance with implementation scripts. All AI-assisted work was reviewed by the authors. AI-generated training data underwent deterministic filtering and checks for completeness, audio quality, and synthesis failures. The authors take full responsibility for the final content of this work, including its text, claims, experiments, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of VoxOPD, we document our methodology, data, and experimental configurations throughout the main paper and appendices. Appendix E describes our training-data construction pipeline in detail. Table 6 reports the source datasets and sample counts, while Figure 6 presents the prompt used to generate reference answers and rationales. The appendix further documents speech-suitability filtering and multi-system speech synthesis. The Method section specifies the key model and algorithmic components, including student-generated trajectories, privileged teacher conditioning, Top-K stop-gradient reverse-KL objectives for text and audio tokens, audio-chunk construction, the shared-credit objective, and the combined training loss. Appendix B provides implementation and training details. Appendix C defines the benchmarks. Appendices A and D describe the protocols for the motivation experiments and human evaluation, respectively.

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.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. *arXiv preprint arXiv:2603.24596*, 2026.

Yifu Chen, Shengpeng Ji, Qian Chen, Tianle Liang, Yangzhuo Li, Ziqing Wang, Wen Wang, Jingyu Lu, Haoxiao Wang, Xueyi Pu, et al. Wavalign: Enhancing intelligence and expressiveness in spoken dialogue models via adaptive hybrid post-training. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2421–2444, 2026a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026b.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Dongjie Fu, Di Cao, Xize Cheng, Zihan Zhang, Wenxu Jia, Yifu Chen, Shengpeng Ji, Yu Zhang, and Tao Jin. X^3-OPD: Distilling reasoning into large audio-language models via on-policy alignment. *arXiv preprint arXiv:2607.21550*, 2026.

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

Jing Hu, Danxiang Zhu, Xianlong Luo, Dan Zhang, Shuwei He, Yishu Lei, Haitao Zheng, Shikun Feng, Jingzhou He, Yu Sun, et al. Cord: Bridging the audio-text reasoning gap via weighted on-policy cross-modal distillation. *arXiv preprint arXiv:2601.16547*, 2026.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 23802–23804, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, et al. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*, 2025.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition, 2020. URL https://arxiv.org/abs/1910.11473.

Minseok Kim, Jingxiang Chen, Seong-Gyun Leem, Yin Huang, Rashi Rungta, Zhicheng Ouyang, Haibin Wu, Surya Teja Appini, Ankur Bansal, Yang Bai, et al. Aligning paralinguistic understanding and generation in speech llms via multi-task reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 636–648, 2026.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. Edge-opd: Internalizing privileged context with evidence guided on-policy distillation. *arXiv preprint arXiv:2605.23493*, 2026.

Jijie Li, Li Du, Hanyu Zhao, Bo wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. Infinity instruct: Scaling instruction selection and synthesis to enhance language models, 2025. URL https://arxiv.org/abs/2506.11116.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6626–6642, 2024.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. Align-slm: Textless spoken language models with reinforcement learning from ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20395–20411, 2025.

Xiang Lin, Tian-Hao Zhang, Chunfeng Wang, Zhou Pan, Kun Zhan, and Liang Li. Towards more expressive spoken llms: Fine-grained intent benchmarking and acoustic-lexical decoupled policy optimization. *arXiv preprint arXiv:2608.03054*, 2026.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, et al. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*, 2025.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*, 2023.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`, 2023.

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.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset, 2024. URL `https://arxiv.org/abs/2402.10176`.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18821–18835, 2026a.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. Blsp-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19186–19199, 2024.

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*, 2026b.

Yuxiang Wang, Qinke Ni, Shengbo Cai, Wan Lin, Liqiang Zhang, and Zhizheng Wu. Parabridge: Bridging paralinguistic perception and dialogue behavior in speech language models. *arXiv preprint arXiv:2606.10581*, 2026c.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL `https://arxiv.org/abs/1707.06209`.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions, 2025. URL https://arxiv.org/abs/2304.12244.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, volume 2026, pp. 70704–70729, 2026.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, et al. Vstyle: A benchmark for voice style adaptation with spoken instructions. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19477–19481. IEEE, 2026.

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.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Wang Benyou, and Haizhou Li. Echomind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *International Conference on Learning Representations*, volume 2026, pp. 55554–55588, 2026.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

A Motivation Experiment Details

Controlled Input Comparison We use Step-Audio 2 mini without additional training and vary only the information provided at inference time. As shown in Figure 1(a), we compare five conditions: speech alone (Audio), the corresponding transcript alone (Text), speech together with its transcript (Audio+Text), the transcript with a reference rationale (Text+Rationale), and speech, transcript, and rationale together (Audio+Text+Rationale). The rationales provide only high-level guidance on the reasoning procedure. They neither state nor paraphrase the target answer, and they do not disclose answer-bearing clues, such as identifying, favoring, or ruling out candidate options.

Audio-Token Perturbation Figure 1(b) studies whether local changes to discrete audio tokens produce commensurate changes in decoded speech. We draw 200 utterances from the Expresso (Nguyen et al., 2023). Sampling is balanced across four speakers and ten speaking styles. We tokenize each utterance using the tokenizer of Step-Audio 2 mini. All token sequences are decoded with the same Token2wav model and the resulting speech is transcribed by Whisper large-v3-turbo (Radford et al., 2023) using greedy English decoding.

For a token sequence of length T and a replacement ratio $r > 0$, we sample without replacement

$$N_r = \max(1, \mathrm{round}(T \times r/100))$$

positions. We use $r \in \{0, 1, 2, 5, 10, 20, 50\}\%$, where $r = 0$ is the unperturbed *codec-clean* baseline, and repeat every nonzero condition three times. At each selected position, we rank codebook entries by squared Euclidean distance between their quantization coordinates and the corresponding pre-quantization continuous representation, excluding the original token. A replacement is then sampled uniformly from the nearest $K_{\mathrm{nn}} \in \{5, 10, 15, 20, 50, 100\}$ entries. For paired comparisons, all neighborhood sizes share the same replacement positions for a given utterance, ratio, and repetition. Before computing WER, we apply Unicode NFKC normalization, uppercase the text, remove punctuation, and collapse repeated whitespace. Let $\hat{y}_0$ denote the ASR transcript of the codec-clean reconstruction and $\hat{y}_r$ that of a perturbed reconstruction. We report ASR-change WER, $\mathrm{WER}(\hat{y}_0, \hat{y}_r)$, which measures perturbation-induced content changes relative to the codec-clean reconstruction while reducing confounding from codec reconstruction and ASR errors.

We additionally measure acoustic similarity between each perturbed reconstruction and its codec-clean counterpart. From both signals, we extract 80-dimensional positive log-mel features $\mathbf{f}_t = \log(1 + 10^4 \mathbf{m}_t)$, align them to their common duration, and retain frames whose codec-clean energy exceeds 10^{-4} of its maximum frame energy. Similarity is the mean framewise cosine similarity over these non-silent frames and is defined as one for the codec-clean condition. For each condition, we first average the three repetitions within each utterance and then compute a macro-average across utterances. We estimate 95% confidence intervals with 5,000 bootstrap resamples, using utterances as clusters.

Modality-Specific Alignment We examine teacher–student alignment throughout training under four objectives: $\mathcal{L}_{\mathrm{text}}$, $\mathcal{L}_{\mathrm{text}} + \mathcal{L}_{\mathrm{audio}}$, $\mathcal{L}_{\mathrm{text}} + \mathcal{L}_{\mathrm{chunk}}$, and the full objective $\mathcal{L}_{\mathrm{text}} + \mathcal{L}_{\mathrm{audio}} + \mathcal{L}_{\mathrm{chunk}}$. We conduct the experiment on 500 randomly sampled examples from an evaluation set disjoint from the training data, covering the same task categories represented in the training corpus. At each evaluation point during training, the student independently samples four on-policy responses to the speech input of each example.

For a sampled response $y = (y_1, \ldots, y_T)$, the student and teacher score the same generated prefix at every position:

$$p_t^S = \pi_\theta(\cdot \mid a, y_{<t}), \qquad p_t^T = \pi_T(\cdot \mid a, x, r, y_{<t}),$$

We aggregate text and audio positions separately. Let $\mathcal{U}_t$ be the intersection of the student and teacher Top-20 token sets at position t, and let $\widetilde{p}_t^S$ and $\widetilde{p}_t^T$ denote their respective distributions renormalized over $\mathcal{U}_t$. The overlap-token advantage is

$$A_t^{\mathrm{overlap}} = \sum_{v \in \mathcal{U}_t} \widetilde{p}_t^S(v) \left[\log \widetilde{p}_t^T(v) - \log \widetilde{p}_t^S(v)\right].$$

This quantity is the negative reverse KL divergence on the shared support, and values closer to zero indicate stronger alignment. We additionally report the full-vocabulary absolute entropy gap,

$$\Delta H_t = \left|H(p_t^S) - H(p_t^T)\right|,$$

for which lower values indicate closer agreement in predictive uncertainty. We first average each metric over tokens of the same modality within a rollout, then over the four rollouts for each example, and finally macro-average over the 500 examples.

B Training Details

During VoxOPD training on Step-Audio 2 mini, the backbone is frozen and only the student LoRA parameters are updated. The LoRA adapters contain 40.37M trainable parameters out of 8.36B total

parameters (0.4832%), with rank 16, scaling factor 32, and dropout 0.05. We apply LoRA to the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules. Training uses `bfloat16` precision and Top-$K = 20$ on $8\times$ NVIDIA H200 GPUs. We use AdamW with a learning rate of 2×10^{-5}, a linear learning-rate schedule, no warm-up, no weight decay, and a maximum gradient norm of 1.0. The per-device batch size is 1, with 4 gradient-accumulation steps, yielding a global batch size of 32.

C Benchmark Details

VoiceBench (Chen et al., 2026b) evaluates whether speech dialogue models can understand spoken instructions and produce semantically appropriate responses. We organize its subsets into two groups. The open-ended dialogue group, denoted VB-Open, contains AlpacaEval, CommonEval, and WildVoice. AlpacaEval assesses response quality on diverse general-purpose instructions. CommonEval targets information-seeking questions recorded by diverse speakers in realistic conditions, while WildVoice, included following the expanded official release, tests open-ended responses to crowdsourced speech with varied speaker accents.

The capability-specific group, denoted VB-Cap, covers six targeted abilities. SD-QA assesses factual question answering from dialectally diverse speech. OpenBookQA targets elementary scientific knowledge and the commonsense needed to apply it, whereas MMSU covers broader knowledge across academic subjects. BBH probes challenging compositional, linguistic, and logical reasoning. IFEval measures adherence to explicit content, structural, and formatting constraints, and AdvBench measures safety alignment when models are presented with harmful requests.

VStyle (Zhan et al., 2026) evaluates voice style adaptation from spoken instructions and therefore focuses on how models use paralinguistic information in generated speech. Acoustic measures explicit control over attributes such as age, gender, speaking rate, pitch, loudness, and emotion, including combinations of multiple attributes. Instruction evaluates open-ended natural-language control of emotion and speaking style, as well as style changes within an utterance. Role Play assesses whether a model can realize vocal traits, emotions, and speaking styles appropriate to a specified scene or character. Empathy tests whether the model can infer an implicitly expressed user state, such as anger, anxiety or fear, sadness or disappointment, or joy or excitement, and respond with supportive content and matching prosody.

D Human Evaluation Protocols

Perceptual effect of audio-token perturbations. We sample 50 Expresso utterances, balanced across four speakers and ten speaking styles. For each utterance, we construct four codec-clean–perturbed pairs, one at each replacement ratio $r \in \{5, 10, 20, 50\}\%$, sampling replacement tokens from the $K_{\mathrm{nn}} = 20$ nearest codebook entries. Trials are shuffled across utterances and replacement ratios rather than presented in increasing order of r, and the clean and perturbed recordings are randomly assigned to the two positions within each pair. Codec-clean–codec-clean control pairs are interspersed with the test pairs. From the three perturbation repetitions available for each utterance and ratio, we select one using a fixed seed, and each resulting pair receives at least five ratings.

Listeners rate overall similarity, content preservation, timbre similarity, prosody similarity, and naturalness on a five-point scale, where 1 indicates the largest difference or lowest quality and 5 indicates no perceptible difference or the highest quality. We report the mean and standard deviation for each dimension and replacement ratio. Higher scores are better.

Small-scale token perturbations are nearly imperceptible, whereas increasing the replacement ratio primarily degrades prosody and naturalness while content remains comparatively robust.

Human evaluation on VStyle. We stratify 40 VStyle items across Acoustic, Instruction, Role Play, Empathy, and their subcategories. The evaluated Step-Audio 2 systems are Base, SFT, GRPO, CORD, ParaBridge, and VoxOPD. For each item, listeners directly compare the responses from all six systems, and each item receives at least five independent evaluations. A trial first plays the user utterance and then the six anonymized responses in random order; response transcripts are not shown, and the original relative loudness is retained.

Table 4: Human evaluation of audio-token perturbations. All dimensions are rated on a 1–5 scale, with higher scores indicating greater perceptual preservation or quality. Values are means ± standard deviations.

r (%)	Overall Similarity	Content	Timbre	Prosody	Naturalness
5	4.97 ± 0.18	4.99 ± 0.11	4.98 ± 0.15	4.93 ± 0.25	4.88 ± 0.33
10	4.67 ± 0.54	4.88 ± 0.36	4.73 ± 0.51	4.59 ± 0.60	4.52 ± 0.66
20	4.28 ± 0.92	4.60 ± 0.63	4.44 ± 0.82	4.12 ± 0.91	4.02 ± 0.92
50	3.49 ± 1.33	4.09 ± 1.02	3.69 ± 1.27	3.27 ± 1.19	3.09 ± 1.24

Table 5: Six-system human evaluation on VStyle items. Scores are means ± standard deviations on a 1–5 scale. The left group reports the mean of the four dimensions by VStyle category, and the right reports dimension scores over all items. Each item receives at least five ratings.

System	Mean overall score (1–5)					Mean dimension score (1–5)			
	Acous.	Instr.	Role	Emp.	Overall	Style	Expr.	Natural.	Content
Base	3.83 ± 0.70	3.30 ± 0.63	3.64 ± 0.66	3.74 ± 0.57	3.62 ± 0.67	3.31 ± 0.97	3.17 ± 0.82	4.21 ± 0.81	3.81 ± 1.19
SFT	4.54 ± 0.42	4.12 ± 0.58	$\mathbf{4.48 \pm 0.56}$	4.32 ± 0.68	4.37 ± 0.59	3.99 ± 1.05	3.94 ± 0.98	4.76 ± 0.49	4.78 ± 0.55
GRPO	3.85 ± 0.63	3.31 ± 0.43	3.64 ± 0.71	4.00 ± 0.63	3.70 ± 0.66	3.41 ± 1.05	3.21 ± 0.91	4.40 ± 0.67	3.79 ± 1.15
CORD	3.90 ± 0.63	3.49 ± 0.56	3.51 ± 0.85	4.12 ± 0.63	3.76 ± 0.72	3.41 ± 1.04	3.33 ± 0.99	4.36 ± 0.75	3.92 ± 1.16
ParaBridge	4.47 ± 0.43	4.02 ± 0.45	3.15 ± 0.57	3.49 ± 0.86	3.78 ± 0.78	3.40 ± 1.00	3.33 ± 0.94	4.32 ± 0.80	4.09 ± 1.22
VoxOPD	$\mathbf{4.79 \pm 0.31}$	$\mathbf{4.41 \pm 0.56}$	4.46 ± 0.58	$\mathbf{4.42 \pm 0.69}$	$\mathbf{4.52 \pm 0.57}$	$\mathbf{4.18 \pm 1.02}$	$\mathbf{4.14 \pm 1.01}$	$\mathbf{4.89 \pm 0.33}$	$\mathbf{4.86 \pm 0.43}$

For each response, listeners provide separate 1–5 scores for style match, prosodic expressiveness, naturalness, and content appropriateness; for Empathy, style match explicitly refers to adaptation to the user's emotional state. We define the overall score for each response as the arithmetic mean of these four dimension scores rather than eliciting a separate overall rating. The primary outcomes are the mean of the four dimensions for each system, reported across all items and by VStyle category. We additionally report each dimension score across all items.

Consistent with the VStyle results in Table 2, VoxOPD achieves the highest overall human-evaluation score and leads every category except Role Play, where SFT is marginally better.

E Dataset Construction

We construct the training set from two components: curated examples selected from open-source corpora and a small set of examples constructed by us. Both components are processed into speech-oriented interactions with reference. The resulting dataset contains 38,532 English spoken-interaction examples.

We aggregate prompts spanning mathematical and logical reasoning (e.g., Infinity (Li et al., 2025), SlimOrca, MetaMathQA (Yu et al., 2023), and OpenMathInstruct-1 (Toshniwal et al., 2024)), science question answering (e.g., ARC (Clark et al., 2018), PubMedQA (Jin et al., 2019), QASC (Khot et al., 2020), and SciQ (Welbl et al., 2017)), code and general instruction following (e.g., OpenAssistant (Köpf et al., 2023), No Robots (Rajani et al., 2023), and WizardLM (Xu et al., 2025)), creative or open-ended dialogue, and safety prompts from BeaverTails (Ji et al., 2023). We additionally construct expressive-interaction and instruction-following examples to supplement the open-source data.

Regular expressions and deterministic rules remove prompts that rely on interface-dependent content, such as URLs, Markdown markup, JSON-heavy structures, or complex visual layouts. We also require prompts to be predominantly ASCII and no longer than 250 words, which avoids unusually long or difficult-to-verbalize inputs.

The remaining prompts are reviewed by a LLM to determine whether they are self-contained and can be understood and answered through speech alone. Dialogue prompts must provide enough context and a clear objective; instruction-following prompts must contain objective, verifiable constraints; logic and mathematics prompts must avoid spatial layouts and use speakable expressions; science prompts must target well-defined facts or principles; and creative prompts must be suitable for bounded spoken responses. During the same review, the LLM assigns each prompt to a task

category and rewrites visually dependent mathematical or logical expressions into a speakable form when needed.

Gemini-3-Flash generates a concise, speech-suitable reference answer and a task-specific rationale for each example. Figure 6 shows the generation prompt. The resulting training set comprises 38,532 examples, with a detailed breakdown by task and source provided in Table 6.

Each text example is converted into a pair of speech utterances: the user prompt and its reference response. We use five TTS systems: CosyVoice3, IndexTTS 2.5, CosyVoice2, Qwen3-TTS, and VoxCPM2. Using multiple synthesizers prevents the training data from inheriting the characteristics of any single system. Reference-speaker characteristics and acoustic prompts are sampled from Seed-TTS-Eval, introducing variation in speaker identity and delivery.

For expressive-interaction examples, the requested tone, role, emotion, or acoustic attribute is converted into a short delivery instruction for response synthesis. This separates what the response should say from how it should be spoken. Before tokenization, we check the synthesized prompt–response pairs for completeness, audio quality, and difficulty-related synthesis failures. Only complete pairs that pass these checks are converted into audio tokens for training.

Table 6: Detailed composition of the final training set.

Task	Source	Quantity
Code	Infinity	1,744
	OpenAssistant	4
Creative	Infinity	618
	Airoboros-create	213
Dialogue	OASST1	50
Instruction	WizardLM	109
	No Robots	102
	Constructed	4,000
Logic	SlimOrca	2,205
Math	Infinity	4,083
	MetaMathQA	9,766
	OpenMathInstruct-1	338
Science QA	ARC-Challenge	621
	ARC-Easy	906
	PubMedQA	4,760
	QASC	2,856
	SciQ	3,006
Open-ended	Capybara	46
	Pure-Dove	20
	Chatbot Arena	4
	UltraChat	81
Safety	BeaverTails	1,000
Expressive Interaction	Constructed	2,000
Total	–	**38,532**

Prompt: Teacher Reference Answer and Rationale Construction

System prompt

```
You generate high-quality teacher privileged inputs for data construction.

For each sample, create:
1. answer_text: a high-quality reference answer a* that can be sent directly to TTS.
2. rationale: a concise reference reasoning / response plan r* that explains how to reach
the answer.

Use the supplied context/evidence when present. Do not invent facts that conflict with the
evidence.

answer_text:
- Write a normal spoken-style answer that is accurate, concise, and easy to read aloud.
- For style-conditioned examples, provide only the words the assistant should say and make
the content compatible with the requested speaking style, role, emotion, or acoustic
attribute.
- Do not describe or repeat the requested speaking style in answer_text unless those words
are part of the intended spoken content.
- Do not write a long explanation in answer_text. Give only the final answer or a brief
direct response.
- Keep answer_text under 80 words. Prefer one sentence. Use two short sentences only if
needed.
- Even for open-ended, creative, dialogue, coding, or advice tasks, provide a concise
TTS-friendly reference answer instead of a long article, list, song, script, or full essay.
- Do not use emoji, Markdown, bold markers such as **, bullet symbols, decorative
separators, or special formatting.
- Do not use parentheses or bracketed asides in answer_text unless they are absolutely
necessary for correctness.
- Avoid text that requires cleanup before speech synthesis, such as code fences, tables,
URLs unless essential, citation markers, stray symbols, or dense mathematical notation.
- For multiple-choice questions, include the option letter and a short natural answer, for
example: "B. Promising."
- Put any explanation or reasoning in rationale, not in answer_text.

rationale:
- Write a concise, high-quality reasoning trace that supports the answer.
- Do not include, quote, or restate the final answer or answer_text in rationale; provide
only the reasoning or response plan.
- Include the key evidence, calculation, rule, or response plan needed to verify the
answer.
- Keep it short and focused. Do not write a long or repetitive chain of thought.
- For context-based questions, explicitly connect the relevant context evidence to the
final answer.
- For math or logic questions, include the essential steps and final check, but avoid
unnecessary derivations.
- For open-ended writing or dialogue tasks, describe the response plan, constraints, and
why the answer satisfies the user request.
- For style-conditioned examples, provide a brief public reasoning summary that identifies
the key content and style constraints the answer is intended to satisfy, including the
requested role, emotion, or delivery characteristics when applicable.
- Do not include hidden uncertainty, filler phrases, or unsupported claims.

Return only a valid JSON object with these keys:
answer_text, rationale, answer_type, confidence.
```

User prompt template

```
Dataset: {dataset_id}
Source JSONL: {source_jsonl}
Source line number: {line_no}
User question (x_text):
{user_question}
Additional evidence/context, when present: {context}
Return a compact JSON object only. Do not wrap it in Markdown.
For multiple-choice questions, answer_text should include the option letter when
appropriate.
Keep rationale concise but sufficient for Scheme A r*.
```

Figure 6: Prompt used to construct reference answers and rationales.

Semantic Reasoning

Example 1

Question: 1000 men have provisions for 18 days. If 400 more men join them, for how many days will the provisions last now?

Rationale: First, calculate the total units of provisions: 1000 men multiplied by 18 days equals 18,000 man-days. After 400 more men join, the total number of men is 1400. To find the number of days, divide the total provisions by the new number of men.

Example 2

Question: Goldenrod and No Hope are in a horse race with 8 contestants. How many different arrangements of finishes are there if No Hope always finishes before Goldenrod and if all of the horses finish the race? Answer Choices: (A) 40320 (B) 20160 (C) 12000 (D) 2400 (E) 2100

Rationale: There are 8 total contestants, and the number of ways they can finish a race is 8 factorial. For No Hope and Goldenrod, there are two possible relative finishing orders. Since these are equally likely across all permutations, exactly half of the arrangements place No Hope before Goldenrod.

Example 3

Question: A crate of apples contains 1 bruised apple for every 30 apples. Three out of every 4 bruised apples are considered not fit to sell, and every apple that is not fit to sell is bruised. If there are 15 apples not fit to sell in the crate, how many apples are there in the crate? Answer Choices: (A) 270 (B) 360 (C) 480 (D) 600 (E) 840

Rationale: First, find the number of bruised apples. Since 3 out of every 4 bruised apples are unfit and there are 15 unfit apples, divide 15 by 3/4. Then use the ratio of 1 bruised apple for every 30 apples to determine the total size of the crate.

Paralinguistic Response

Example 1

Question: I'm sitting in the waiting room for a high-stakes job interview and I can feel my palms getting sweaty and my breathing getting shallow. Could you talk to me in a gentle, reassuring voice to help me manage this anxiety?

Rationale: The response provides a gentle and reassuring message to help the user manage interview anxiety. It focuses on validating their feelings and encouraging slow breathing while maintaining a supportive tone.

Example 2

Question: Listen, I need to vent. I just found out that I was passed over for a promotion I worked so hard for, and the person who got it doesn't even have half my experience. I'm absolutely furious. Talk to me like a close friend who understands why I'm so upset.

Rationale: The response adopts an empathetic tone to validate the user's feelings of anger and disappointment. It focuses on acknowledging their hard work and the unfairness of the situation as a supportive friend would.

Example 3

Question: Explain why you are upset that someone broke your favorite piece of equipment after you specifically asked them not to touch it.

Rationale: The response expresses intense anger by focusing on the disregard for a direct instruction and the resulting damage. The language is direct and sharp to match the requested emotional tone.

Figure 7: Examples of questions and their corresponding rationales used in data construction.