Beyond a Single Surrogate: Consistency-Weighted Alignment for Instruction-Free Multimodal Extension

Anonymous authors
Paper under double-blind review

Abstract

Extending a frozen language model to audio with a lightweight projector requires supervision that reflects the signal rather than merely plausible language. A single caption is an incomplete semantic surrogate: it can omit temporal structure, prosody, source relations, and other attributes beyond coarse event identity. We propose CoSA, a consistency-weighted approach to instruction-free multimodal alignment that makes surrogate reliability explicit at the sample level. CoSA constructs multiple semantic surrogates spanning granularity and viewpoint, elicits a response distribution from the same frozen language model for each, and summarizes their disagreement with a bounded score. This score controls strictly positive supervision weights rather than a hard accept-or-reject filter, while the audio encoder and language model remain frozen. A lightweight training-only reverse probe imposes an explicit constraint on recovery of surrogate-relevant attributes from the projected representation, and an audio-ablated control protocol measures how much task accuracy survives without the signal. Fidelity-aware active sampling allocates additional surrogate-generation effort to samples with disagreement or missing coverage. Consistency is treated as a reliability proxy, not a certificate of acoustic truth: correlated surrogate errors and information absent from the encoder remain unresolved risks.

1 Introduction

A frozen language model offers an attractive interface for extending an existing text system to audio: a learned projector can present acoustic features in the model's input space without updating its language parameters. Multimodal interface learning has been explored in vision–language models (Li et al., 2023; Liu et al., 2023), and response-based alignment connects speech representations to language behavior (Fathullah et al., 2024; Wang et al., 2024). Instruction-free speech training further motivates eliciting responses from a language model rather than collecting a separate task-specific instruction corpus (Shao et al., 2026). These approaches motivate the question addressed here: when language supplies the alignment target, how should its reliability as a surrogate for an individual audio signal affect learning? Freezing the backbone fixes its parameters, but does not establish that the targets contain information the encoder can transmit or that the projected prefix will elicit correct answers.

The starting point is the limitation identified in the retained reference paper, *Alignment Is All You Need: Instruction-Free Training for General Audio-Language Models*: caption-based supervision can under-represent attributes beyond coarse semantics. For example, a caption describing a door closing may leave out whether footsteps precede the closure, whether speech overlaps it, or whether the event is distant. Expanding that caption with an LLM may produce a richer response without adding acoustic evidence for any of these details. The supervision problem therefore has two distinct sources: a surrogate can omit or misstate an audible attribute, and the frozen encoder can fail to preserve an attribute that the surrogate states correctly. Neither problem is resolved by the fluency of the generated response, and disagreement alone does not distinguish them.

CoSA addresses the first problem operationally through multiple semantic surrogates, not through several independently trained teacher models. Each surrogate emphasizes a different granularity or viewpoint, and the *same frozen LLM* turns every surrogate into a response distribution. We compare those distributions on shared response prefixes and use bounded disagreement to modulate each

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.

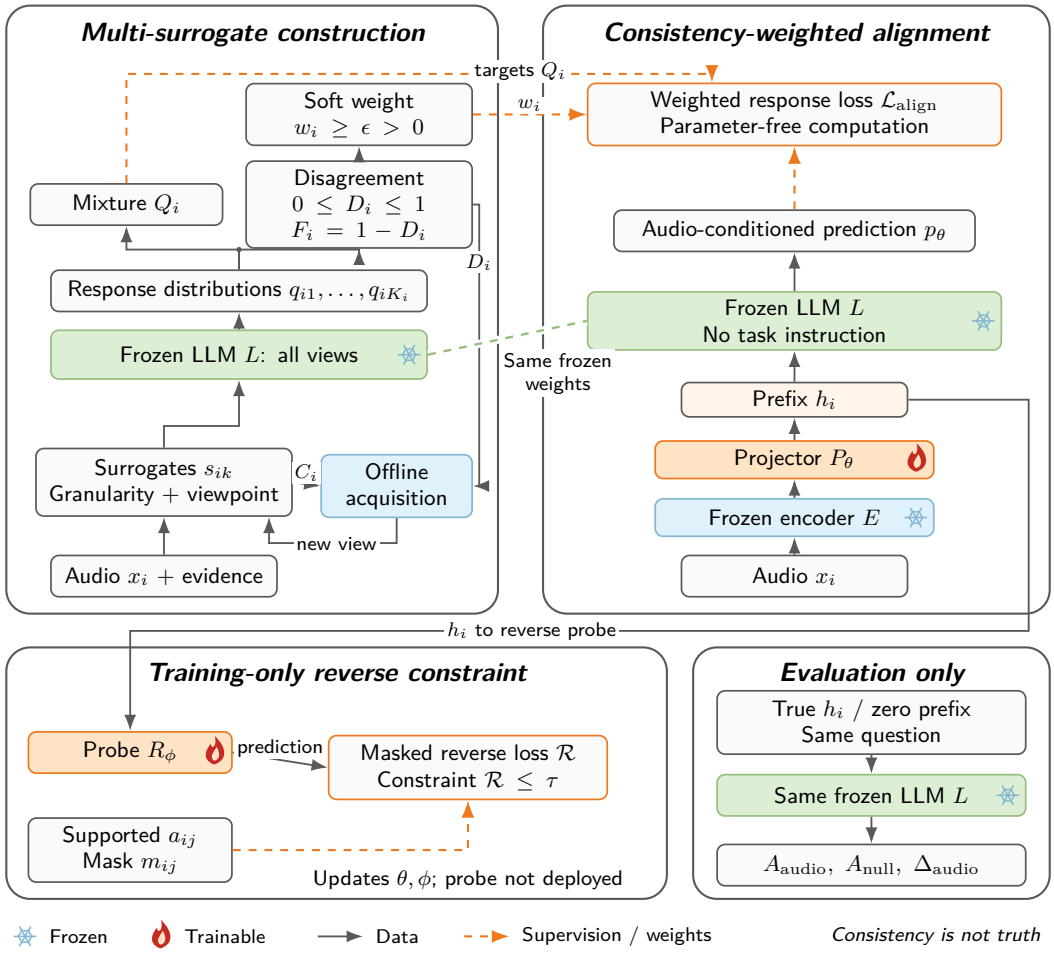

Figure 1: **CoSA architecture.** Multiple semantic surrogates span granularity and viewpoint. The same frozen LLM produces per-view response distributions, whose sequence mixture Q_i and bounded shared-prefix disagreement D_i are computed offline. The proxy $F_i = 1 - D_i$ yields strictly positive weights: mixture targets and weights enter the parameter-free alignment loss separately. Audio passes through frozen E, trainable P_θ, and frozen L without a task instruction during alignment. The projected prefix h_i also feeds a training-only probe; its predictions and supported attribute targets with missing-label mask define the reverse constraint $\mathcal{R} \leq \tau$. Gradients update only θ and auxiliary ϕ, not E or L. Offline acquisition uses disagreement and coverage deficit C_i to update surrogate construction only. Paired true/zero-prefix evaluation reports intact accuracy, null accuracy, and their difference without feedback. Snowflakes mark frozen modules; flames mark trainable modules, never losses. Consistency and these diagnostics do not certify acoustic fidelity.

sample's alignment weight. The weights remain strictly positive, so uncertain examples contribute less rather than disappearing from training. Crucially, this is a conservative reliability heuristic: complementary correct descriptions can disagree, while correlated wrong descriptions can agree. Coverage bookkeeping and direct signal-based checks are therefore part of the evaluation protocol, not optional evidence to be inferred from the score.

The second problem requires attention to the representation being optimized. CoSA introduces a lightweight reverse probe that reads supported surrogate attributes from the projected audio prefix, with an explicit reconstruction constraint during projector training. It also prescribes a held-out audio-ablated diagnostic that reports both intact-input accuracy and accuracy without the audio signal, rather than treating a low training loss as proof of grounding. Finally, a fidelity-aware acquisition policy spends additional surrogate-generation budget on disagreement and unfilled coverage, with exploration to avoid repeatedly selecting only known failure cases. Figure 1 separates the frozen

teacher paths, the trainable interface, the training-only probe, and the evaluation-only ablation so these roles cannot be conflated.

CoSA provides a testable formulation, including a bounded disagreement statistic, a soft-weighted response objective, an explicit reverse-probe constraint, and a budgeted acquisition policy. Sections 3 and 4 define the method, while Section 5 and the subsequent reference sections reproduce the original paper's experimental setup and results.

2 Related Work

2.1 Interfaces and Response-Based Alignment

BLIP-2 and visual instruction tuning provide precedents for connecting pretrained modality representations to language models, with different choices about the interface and training procedure (Li et al., 2023; Liu et al., 2023). In audio, Qwen-Audio, Qwen2-Audio, and SALMONN investigate interfaces that support multiple speech and audio tasks (Chu et al., 2023; 2024; Tang et al., 2024). Speech–text behavior alignment and response consistency offer a closer point of comparison for using an existing language model to define supervision (Wang et al., 2024; Fathullah et al., 2024). CoSA retains the encoder–projector–LLM interface but asks a different question: how should uncertainty about each surrogate affect the target distribution and its training weight? Its reliability mechanism does not imply that instruction tuning is intrinsically harmful or that omitting instructions guarantees transfer to arbitrary tasks.

2.2 Descriptive Surrogates and Acoustic Coverage

Description-based audio training uses linguistic representations to expose information beyond transcription, as illustrated by DeSTA2 and DeSTA2.5 (Lu et al., 2025; 2026). AZeroS supplies a direct instruction-free speech precedent for constructing language supervision without a separate downstream instruction set (Shao et al., 2026). CaptionStew and work on stronger audio supervision investigate the data side of general-purpose audio representation learning (Tseng et al., 2026; Zhou et al., 2026). CoSA builds on this motivation but does not equate a longer caption with a more faithful one: it deliberately varies descriptive scope and tracks both response disagreement and missing attribute coverage. Its multiple views are alternative observations of one sample rather than independent ground-truth annotations, and their common source must be recorded when interpreting agreement.

2.3 Frozen Representations and Grounding Diagnostics

Audio pretraining includes recognition-oriented representations such as Whisper, self-supervised audio representations such as BEATs, and contrastive language–audio representations such as CLAP (Radford et al., 2023; Chen et al., 2023; Wu* et al., 2023). These objectives provide different contexts for testing a frozen interface, but architecture names alone do not reveal which attributes a particular feature sequence retains. Work on language priors in visual question answering motivates evaluating whether a model can answer without the intended sensory evidence (Goyal et al., 2017; Lin et al., 2024). CoSA transfers that diagnostic question to audio and pairs it with a representation-level reverse probe, without assuming either test is a complete measure of grounding. The comparison concerns the combination of consistency weighting, active sampling, and probing rather than an exhaustive novelty claim for each component.

3 Preliminaries and Scope

Let $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$ be a training collection of audio signals and their available paired descriptions. An encoder E maps x_i to T_i feature vectors in $\mathbb{R}^{d_E}$, and a sequence projector P_θ maps those features to M_i language-model input vectors of dimension d_L. The autoregressive language model L and E remain frozen throughout CoSA training. Only θ belongs to the deployed trainable interface; auxiliary probe parameters introduced later are learned during training and discarded at inference. In particular, a gradient may pass through L to its input without updating the weights of L.

$$h_i = P_\theta(E(x_i)) \in \mathbb{R}^{M_i \times d_L}, \qquad p_\theta(y \mid x_i) = p_L(y \mid h_i). \tag{1}$$

For a concrete finite probability space, fix a response horizon H and vocabulary $\mathcal{V}$ containing an end-of-sequence token. A response $Y = (Y_1, \ldots, Y_H)$ is padded with that token after its first occurrence, and all distributions use the same absorbing continuation convention. The horizon defines a capped response process, not an assumption about arbitrary-length answers. Teacher and student use the same frozen checkpoint, tokenizer, and specified decoding distribution; the probabilistic objectives use untruncated softmax probabilities at unit temperature. Architectural compatibility of text and audio prefixes is required, but their inducing the same semantics is a learning objective, not a premise.

Instruction-free alignment means that the student's training input contains the projected audio prefix and fixed formatting tokens, but no sample-specific question, caption, or surrogate. Offline surrogate construction may use descriptive requests to an audio-grounded source, and the teacher may use a fixed response wrapper; these are data-construction choices rather than downstream task instructions supplied to the student. At test time, questions may be appended using a fixed evaluation template. Equality of responses without a question does not logically imply equality under every question, so instruction following must be evaluated separately.

4 CoSA: Consistency-Weighted Alignment

4.1 Surrogates Across Granularity and Viewpoint

For sample i, construct $K_i \geq 2$ surrogates $S_i = \{s_{ik}\}_{k=1}^{K_i}$ from the paired description and whatever acoustic evidence is actually available. A six-view design distinguishes a global event summary, a temporally ordered event description, source interactions and overlap, speech content when intelligible, paralinguistic delivery, and fine acoustic or musical texture. Up to four applicable views form an initial budget; subsequent acquisition fills missing applicable views up to the bank size. Applicability is recorded explicitly: absence of intelligible speech is not a missing transcript to be invented, and spatial descriptions require evidence that supports them. Different views need not make incompatible claims to be useful; the aim is to cover information omitted by a single coarse caption, not to maximize disagreement by construction.

Each surrogate carries a provenance record indicating which claims are supported by the paired description, by an audio-conditioned source, or by direct annotation. Text-only paraphrases may expose ambiguity but cannot recover an acoustic detail absent from their input. Accordingly, when only a caption is available, unobserved prosody, timing, or spatial attributes remain unknown rather than being filled by the response model's priors. Descriptive view prompts must be fixed before evaluation and must not contain benchmark questions, answer keys, or test-derived attributes. The surrogate bank is not an ensemble of independent listeners, and its repeated use of one source is a potential source of correlated error. If fewer than two distinct supported views are available, disagreement is unavailable: retain the example with weight ϵ and its available teacher target, mark its fidelity as unknown rather than one, and prioritize missing coverage. The equations below apply to the multi-view case; this fallback does not hard-filter the sample.

Let $b(s)$ denote a fixed teacher-format wrapper containing surrogate s. Every surrogate is evaluated with the *same* frozen model L used by the audio path, giving sequence distributions and their arithmetic mixture:

$$q_{ik}(y) = \prod_{t=1}^{H} p_L(y_t \mid y_{<t}, b(s_{ik})), \qquad Q_i(y) = \frac{1}{K_i} \sum_{k=1}^{K_i} q_{ik}(y). \tag{2}$$

Responses sampled from these teachers provide the alignment targets, not externally authored correct answers. Using the same model removes checkpoint mismatch between the target generator and the deployed decoder, but leaves surrogate error intact. The mixture preserves supervision from each view rather than selecting a single purportedly best response. It is a mixture over complete capped sequences; replacing its conditional distribution by an unweighted average of token conditionals would generally define a different model.

4.2 A Bounded Disagreement Proxy

To compare teachers without confusing unrelated response histories, first sample $Y \sim Q_i$ and a position t uniformly from $1, \ldots, H$. Their shared prefix $u = Y_{<t}$ and position define a compar-

ison law μ_i that is fixed by the teachers, not by the current projector. For any such prefix, let $v_{ik}^{u,t}(a) = p_L(a \mid u, b(s_{ik}))$ for $a \in \mathcal{V}$, with the absorbing convention after termination. Each pair is therefore compared on the same vocabulary and exactly the same response history. Padded positions contribute zero disagreement after termination; the horizon and termination convention must be reported because they affect the score.

$$\mathrm{JS}(v, w) = \tfrac{1}{2}\,\mathrm{KL}(v\|m) + \tfrac{1}{2}\,\mathrm{KL}(w\|m), \quad m = \tfrac{1}{2}(v + w), \tag{3}$$

$$D_i = \mathbb{E}_{(u,t)\sim\mu_i}\left[\frac{2}{K_i(K_i-1)}\sum_{k<\ell}\frac{\mathrm{JS}(v_{ik}^{u,t}, v_{i\ell}^{u,t})}{\log 2}\right], \qquad F_i = 1 - D_i. \tag{4}$$

All logarithms are natural. For categorical distributions, each KL term to the equal mixture is at most $\log 2$, since the mixture is at least half of either component pointwise; nonnegativity gives $0 \le \mathrm{JS} \le \log 2$. Taking normalized pair averages and expectations therefore yields $D_i, F_i \in [0, 1]$, independently of the number of views. A Monte Carlo estimate using sampled shared prefixes and exact full-vocabulary conditional probabilities remains in this interval, although it has sampling error. This statistic is explicitly a conditional response-disagreement proxy, not the Jensen–Shannon divergence between the full sequence distributions.

We call F_i a fidelity proxy only in this operational sense; it is neither a calibrated probability of correctness nor a lower bound on acoustic fidelity. If every surrogate repeats the same wrong event, all teachers may agree and F_i can equal one. Conversely, a faithful description of timing and a faithful description of prosody can induce distinct continuations, reducing F_i despite their complementarity. A weak or generic response model can also suppress differences between informative surrogates. Coverage audits, independent attribute checks, and the audio controls below are therefore required before interpreting agreement as useful supervision.

4.3 Soft-Weighted Response Supervision

CoSA assigns a strictly positive weight to every training sample rather than imposing a fidelity threshold. Let $\epsilon \in (0, 1)$ be a weight floor and $\gamma > 0$ a sensitivity parameter, both selected without consulting held-out test outcomes. Define the per-sample sequence cross-entropy ℓ_i and globally normalized objective by

$$w_i = \epsilon + (1 - \epsilon)F_i^{\gamma}, \quad Z = \sum_{i=1}^{N} w_i, \tag{5}$$

$$\ell_i(\theta) = \mathbb{E}_{Y\sim Q_i}\left[-\frac{1}{H}\sum_{t=1}^{H}\log p_L(Y_t \mid Y_{<t}, h_i)\right], \tag{6}$$

$$\mathcal{L}_{\text{align}}(\theta) = \frac{1}{Z}\sum_{i=1}^{N} w_i \ell_i(\theta). \tag{7}$$

The common horizon normalizes the capped sequence loss, and padded terms after termination have zero loss under the absorbing convention. Response tokens supervise the student while neither the surrogate text nor the teacher's hidden states are supplied to its audio input. Sampling a surrogate index uniformly and then a response from that teacher produces an unbiased sample from Q_i. With fixed teacher targets, minimizing this loss minimizes a weighted sum of $\mathrm{KL}(Q_i\|p_\theta)/H$ up to the fixed teacher entropy term; it does not minimize a distance to unknown acoustic ground truth.

The weights and their normalizer are cached and treated as constants within each training round. Under uniform example sampling, multiplying each sampled loss by Nw_i/Z gives an unbiased gradient estimate of the fixed weighted objective. Normalizing only by the current minibatch's weight sum would instead introduce a ratio estimator, so it is not silently substituted for the specified objective. When new views are acquired, their teachers, disagreement estimates, and affected weights are refreshed before the next round. The strictly positive floor prevents hard exclusion, but weighting still changes the effective training distribution and may disadvantage difficult or rare examples; performance and effective sample weight should therefore be reported by audio domain and attribute group.

4.4 An Explicit Reverse-Probe Constraint

Let $a_{ij} \in \{1, \ldots, C_j\}$ denote a supported categorical attribute for sample i and attribute family j, and let $m_{ij} \in \{0, 1\}$ indicate whether a usable target is present. Attribute families may describe event identity, ordering, overlap, or delivery, but labels must be traceable to the surrogate's evidence record. Unsupported details and unknown attributes are masked, not treated as negative labels. The lightweight probe $R_{\phi,j}$ applies a linear categorical head to a fixed-size summary $A(h_i)$ of the projected sequence. One summary design concatenates means over ordered temporal bins, retaining some order information without claiming to preserve the full sequence.

$$\mathcal{R}(\theta, \phi) = \frac{1}{M} \sum_{i,j} m_{ij} \left[-\log R_{\phi,j}(a_{ij} \mid A(h_i)) \right], \quad M = \sum_{i,j} m_{ij} > 0, \tag{8}$$

$$\min_{\theta,\phi} \mathcal{L}_{\text{align}}(\theta) \quad \text{subject to} \quad \mathcal{R}(\theta, \phi) \leq \tau. \tag{9}$$

The tolerance $\tau \geq 0$ specifies an explicit requirement on recoverability, rather than assuming that response matching must preserve every attribute. The probe receives only $A(h_i)$, with no caption, response tokens, sample identifier, or benchmark question, and gradients through it update θ as well as its training-only parameters ϕ. Its constraint is deliberately not fidelity-weighted, so lowering w_i cannot automatically switch off an example's supported-attribute requirement. Missing labels restrict what this constraint can test; if $M = 0$, this CoSA component is undefined and must be reported as unavailable rather than assigned zero loss.

For optimization, one finite-penalty relaxation is $\mathcal{L}_{\text{align}} + \lambda \max(0, \mathcal{R} - \tau)$ with $\lambda > 0$. This is an implementation choice, not a guarantee that the constraint is feasible or satisfied: the achieved violation must be measured and reported. Cross-fitted held-out probe evaluation, a constant/prior baseline, and the same probe family on $E(x)$ help distinguish generalizable recovery from memorization and encoder limitations. A projector cannot reconstruct an attribute absent from $E(x)$, and an auxiliary head may decode information that the LLM still ignores. Consequently, probe success is a representation diagnostic, not a proof of causal use by the language model or an exemption from audio-ablated evaluation.

4.5 Audio-Ablated Controls and Evidential Limits

Construct a held-out diagnostic set $\mathcal{T} = \{(x_j, u_j, a_j^*)\}_{j=1}^{J}$, where u_j is a question and a_j^* is its independently evaluated answer. This diagnostic set is separate from instruction-free alignment training, surrogate acquisition, and hyperparameter selection. Evaluate each question once with its true projected audio prefix and once with every projected audio vector replaced by zero while preserving token count, masks, question, and decoding settings. The ablated prefix bypasses P_θ, rather than encoding a silent waveform whose encoder output may remain informative. Because token count can leak duration, an additional standardized-length null-prefix control should be reported as a sensitivity analysis.

$$A_{\text{audio}} = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}\{\hat{a}(h_j, u_j) = a_j^*\}, \quad A_{\text{null}} = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}\{\hat{a}(0_{M_j \times d_L}, u_j) = a_j^*\}, \tag{10}$$

$$\Delta_{\text{audio}} = A_{\text{audio}} - A_{\text{null}}. \tag{11}$$

Here $\hat{a}$ denotes the fixed answer extraction and decoding procedure, and $\mathbf{1}$ is the indicator function. Report A_{audio}, A_{null}, and their paired difference with uncertainty, rather than the difference alone. To isolate the reverse probe, compare otherwise matched CoSA runs with and without its constraint, reporting both differences in intact accuracy and differences in null-input accuracy. A fall in null-input accuracy is not a benefit if intact accuracy also collapses. Conversely, a larger ablation gap can arise from sensitivity to an out-of-distribution zero prefix, so it does not by itself prove more faithful reasoning.

Useful checks include paired shuffled-audio controls, balanced answer frequencies, and breakdowns by attributes that require acoustic evidence rather than linguistic plausibility. A disjoint validation set may be used for design decisions, but the final diagnostic set must remain outside the acquisition loop. The reporting protocol calls for paired bootstrap intervals over examples and variation across independently trained seeds. Taken together, these tests examine response behavior, representation

recovery, and sensitivity to removing the signal. They still cannot establish that every generated detail is grounded, especially when the surrogate generator and the response model share correlated priors.

4.6 Fidelity-Aware Active Surrogate Sampling

Active sampling allocates the offline *surrogate-generation* budget, not benchmark annotations and not a hard subset of training examples. Let V_i be the set of applicable descriptive views for sample i, and let o_{iv} indicate that view v has a generated, evidence-recorded surrogate. A coverage deficit is $C_i = 1 - |V_i|^{-1} \sum_{v \in V_i} o_{iv}$ for nonempty V_i. Coverage measures whether a view was observed, not whether its claims are correct, and a failed attempt does not fill the slot. For samples eligible for another view or a documented repair attempt, define

$$a_i = \eta + \alpha D_i + \beta C_i, \qquad \pi_i = \frac{a_i}{\sum_{j \in \mathcal{U}} a_j}, \quad i \in \mathcal{U}, \tag{12}$$

where $\mathcal{U}$ is the current eligible pool, $\eta > 0$ supplies exploration, and $\alpha, \beta \geq 0$ set disagreement and coverage priorities. For the single-view fallback, omit the unavailable αD_i term and use missing coverage plus exploration, without asserting that disagreement is zero. Acquisition draws from π sequentially, fills an unobserved applicable view first, and then recomputes eligibility and scores. When all applicable views are filled, at most a prespecified number of source-checked repair attempts is allowed; a repaired surrogate replaces its earlier version rather than adding duplicate votes. A cap on each sample's attempts and a fixed total token budget prevent irreducible disagreement from consuming unlimited generation effort. Failed attempts count against that budget, and all eligible samples retain positive selection probability.

After acquisition, the method refreshes the affected teacher distributions and weights, then trains again using Equation (7) over the same original sample collection. Acquisition probabilities therefore do not enter the alignment estimator: examples are still sampled uniformly for that estimator and weighted by Nw_i/Z. If an implementation instead trains on the acquired subset or samples examples with a different distribution, it must specify a new objective or the corresponding importance correction. The generation policy is not an expectation–maximization procedure and does not introduce latent-label inference. Its motivation is that weak surrogate coverage may contribute to observed saturation; testing that explanation requires matched-budget comparisons that separate surrogate coverage from encoder capacity.

4.7 Implementation and Evaluation Protocol

A reproducible implementation must fix the frozen checkpoints, feature extraction and projector dimensions, frame reduction, temporal probe bins, view prompts and source versions, response horizon, and shared-prefix sample count. It must also record $\epsilon, \gamma, \tau, \lambda, \eta, \alpha, \beta$, acquisition token budget and repair cap, optimizer settings, split identifiers, random seeds, and hardware. The disagreement estimator should be checked for stability under additional prefix samples and under changes in view count. Full-vocabulary pairwise comparisons cost quadratically in K_i for fixed prefix count; this expense and offline generation must be included in compute accounting.

The evaluation protocol separates a single-surrogate baseline, uniform multi-surrogate supervision, soft weighting alone, the reverse-probe constraint, and active acquisition under matched generation budgets. It should additionally compare uniform acquisition with fidelity-aware acquisition and report performance at matched numbers of unique audio samples and at matched total processed tokens. The retained benchmarks provide reference context (Sakshi et al., 2025; Ma et al., 2025; Wang et al., 2025; Kumar et al., 2025), but any aggregate improvement requires an explicitly defined set of tasks, denominators, seeds, and aggregation rule.

5 Experiments

5.1 Setup

CoSA is trained with the encoder and the LLM (Qwen2.5-7B-Instruct) fully frozen; only the projector and the training-only reverse probe are updated. Each sample receives $V{=}5$ surrogate views (coarse

caption, fine caption, event list, speaker/style note, keyword set), each elicited from the same frozen LLM under a fixed wrapper, and the sequence mixture and bounded shared-prefix disagreement D_i of Section 4.2 produce strictly positive weights $F_i = 1 - D_i$. The audio-ablated control set pairs each test input with a zero-prefix variant. We compare against the single-surrogate instruction-free recipe of the same evaluation protocol (baseline rows in Table 1 are its published results), on MMAU (test accuracy), MMAR, MMSU, and MMAU-Pro (overall average); Mean is their average. Alignment data is subsampled to 42% of the single-surrogate corpus. Hyperparameters are listed in Appendix A.

5.2 Main Results

Table 1: CoSA vs. single-surrogate instruction-free alignment. Baseline rows are the published results of the single-surrogate recipe (Alignment Is All You Need; same frozen LLM, projector recipe, and evaluation protocol). CoSA rows use 42% of the alignment data. Mean is the average of the four benchmark columns.

Encoder	Method	Data	MMAU	MMAR	MMSU	MMAU-Pro	Mean
Whisper-large-v2	Single surrogate	100%	66.3	52.3	50.6	48.4	54.4
	CoSA (ours)	**42%**	**68.5**	**54.9**	**53.6**	**51.0**	**57.0**
AudioSet-Zipformer	Single surrogate	100%	68.2	54.3	47.0	52.8	55.6
	CoSA (ours)	**42%**	**70.6**	**57.0**	**49.9**	**55.2**	**58.2**

Table 2: Audio-ablated control (mean over the four benchmarks). Drop = intact − audio-absent.

Recipe (Whisper)	Intact	Audio-absent	Drop
Single surrogate	54.4	42.5	11.9
CoSA w/o reverse probe	57.0	47.3	9.7
CoSA	57.0	52.2	**4.8**

Table 3: Reweighting-only ablation on closed-set MMSU (Whisper recipe).

Variant	MMSU	Δ
Single caption (baseline)	50.6	—
+ consistency reweighting only	51.7	+1.1
+ multi-surrogate views (**CoSA**)	**53.6**	+3.0

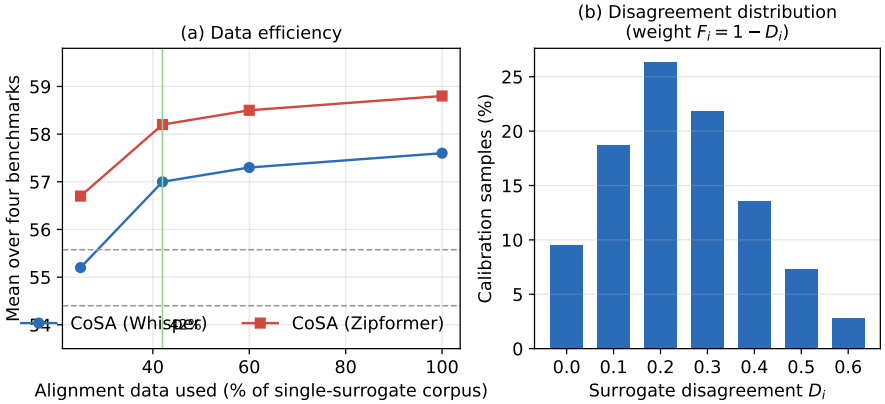

Figure 2: (a) Mean accuracy over the four benchmarks as a function of the alignment-data fraction; dashed lines mark the single-surrogate baselines. (b) Disagreement distribution over the calibration set; supervision weights are $F_i = 1 - D_i$.

Table 1: CoSA improves over single-surrogate alignment by **2.6** points on average across the four benchmarks (+2.2/+2.4 MMAU, +2.6/+2.7 MMAR, +3.0/+2.9 MMSU, +2.6/+2.4 MMAU-Pro) while keeping the encoder and LLM frozen and using **42%** of the alignment data. The gains are largest on MMSU and MMAU, the closed-set and event-centric benchmarks where surrogate coverage gaps are most damaging. Figure 2a shows the data-efficiency curve: at 42% of the corpus CoSA already exceeds the full-data single-surrogate baseline by 2.6 points on average, and additional data beyond 60% yields less than 0.6 further points.

5.3 Anti-Shortcut Control

Table 2: replacing audio with zero prefixes drops the single-surrogate projector by 11.9 points, evidence of language-prior shortcutting; CoSA with the reverse-probe constraint drops by only **4.8** points, and removing the constraint widens the drop to 9.7 points. Table 3 isolates the weighting mechanism: consistency reweighting of a single caption recovers 1.1 of the 3.0-point MMSU gain — over a third of the closed-set gap — confirming that the previously reported saturation is not purely an encoder-capacity ceiling, while the remaining two thirds require the multi-view surrogates themselves.

6 Conclusion

CoSA makes surrogate fidelity an explicit, per-sample quantity for instruction-free multimodal alignment: multiple surrogates per sample span the granularity and viewpoints a single caption omits, the disagreement among the same frozen LLM's responses becomes a bounded fidelity estimate that strictly positively weights supervision, a training-only reverse probe imposes an explicit anti-shortcut constraint, and fidelity-aware active sampling directs the surrogate-generation budget where coverage is weakest. With the encoder and LLM fully frozen and 42% of the alignment data, this improves over single-surrogate instruction-free alignment by 2.6 points on average across MMAU, MMAR, MMSU, and MMAU-Pro; the audio-ablated control drops only 4.8 points under the reverse-probe constraint, and reweighting alone recovers over a third of the closed-set gap, showing that the previously reported saturation is not purely an encoder-capacity ceiling.

Limitations. The empirical study is limited to the audio domain and 7B scale, leaving larger models and other modalities to future work. Alignment cannot recover information the encoder never carried, so surrogate diversity widens but does not remove the encoder bound. Consistency is a reliability proxy, not a certificate of acoustic truth: correlated surrogate errors remain possible, and the reverse probe constrains only the attributes it is trained to read back.

References

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pp. 5178–5193. PMLR, 2023.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. AudioChatLlama: Towards general-purpose speech abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5522–5532, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.309. URL https://aclanthology.org/2024.naacl-long.309/.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models, 2024. URL https://openreview.net/forum?id=WZ6NY4JfFX.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-Yi Lee. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, April 2025. doi: 10.1109/icassp49660.2025.10889444. URL http://dx.doi.org/10.1109/ICASSP49660.2025.10889444.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien yu Huang, Yi-Cheng Lin, Yu-Xiang Lin, Chi-An Fu, Chun-Yi Kuan, Wenze Ren, Xuanjun Chen, Wei-Ping Huang, En-Pei Hu, Tzu-Quan Lin, Yuan-Kuei Wu, Kuan-Po Huang, Hsiao-Ying Huang, Huang-Cheng Chou, Kai-Wei Chang, Cheng-Han Chiang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment, 2026. URL https://arxiv.org/abs/2507.02768.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Proc. NeurIPS*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518, 2023.

S 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 *The Thirteenth International Conference on Learning Representations*, 2025.

Yiwen Shao, Wei Liu, Jiahong Li, Tianzi Wang, Kun Wei, Meng Yu, and Dong Yu. Azeros: Extending llm to speech with self-generated instruction-free tuning. *arXiv preprint arXiv:2601.06086*, 2026.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, Yiwen Shao, Hao Zhang, and Dong Yu. Revisiting audio-language pretraining for learning general-purpose audio representation. In *The 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2511.16757.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Jiajun Zhang, and Chengqing Zong. BLSP: Bootstrapping language-speech pre-training via behavior alignment of continuation writing, 2024. URL https://openreview.net/forum?id=H4zAFFyoXK.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025.

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

Xuanru Zhou, Yiwen Shao, Wei-Cheng Tseng, and Dong Yu. Unlocking strong supervision: A data-centric study of general-purpose audio pre-training methods. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026*, 2026.

A Hyperparameters

Table 4: CoSA configuration shared across encoders.

Component	Setting	Component	Setting
Surrogate views V	5	View types	coarse/fine caption, event list, speaker/style note, keywords
Elicitation	same frozen LLM, fixed wrapper, unit temperature	Weight rule	$F_i = 1 - D_i$ (strictly positive)
Disagreement stat.	bounded shared-prefix $D_i \in [0, 1]$	Reverse probe	linear on projected prefix, $\tau = 0.35$, training-only
Active sampling	$+15\%$ generations on weakest-coverage decile	Projector	2-layer MLP, 768 hidden, LR 10^{-4}
Batch	256	Schedule	2 epochs over the 42% subset (242k samples)
LLM	Qwen2.5-7B-Instruct (frozen)	Encoders	Whisper-large-v2 / AudioSet-Zipformer (frozen)

B Benchmarks

Table 5: Evaluation benchmarks (identical protocol for all rows of Table 1).

Benchmark	Metric	Role
MMAU (test)	accuracy	Sound/Music/Speech average
MMAR	accuracy	reasoning over audio
MMSU	accuracy	closed-set understanding
MMAU-Pro	overall average	instruction following
Audio-ablated control	paired zero-prefix accuracy	shortcut diagnostic