Understanding the Modality Gap: An Empirical Study on the Speech-Text Alignment Mechanism of Large Speech Language Models

Anonymous ACL submission

Abstract

End-to-end Large Speech Language Models (LSLMs) have demonstrated impressive conversational generation abilities, yet consistently fall short of traditional pipeline systems on semantic understanding benchmarks. In this work, we reveal through systematic experimentation that although LSLMs lose some text input performance after speech-text alignment training, the performance gap between speech and text inputs is more pronounced, which we refer to as the **modality gap**. To understand this gap, we analyze both coarse- and fine-grained text and speech representations. At the coarse-grained level, representations of speech and text in deeper layers are found to be increasingly aligned in direction (cosine similarity), while concurrently diverging in magnitude (Euclidean distance). We further find that representation similarity is strongly correlated with the modality gap. At the fine-grained level, a spontaneous token-level alignment pattern between text and speech representations is observed. Based on this, we introduce the Alignment Path Score to quantify token-level alignment quality, which exhibits stronger correlation with the modality gap. Building on these insights, we design targeted interventions on critical tokens through angle projection and length normalization. These strategies demonstrate the potential to improve correctness for speech inputs. Our study provides the first systematic empirical analysis of the modality gap and alignment mechanisms in LSLMs, offering both theoretical and methodological guidance for future optimization.

1 Introduction

The emergence of Large Speech Language Models (LSLMs) has revolutionized human-computer interaction by enabling direct processing of both speech representations and text inputs, subsequently generating textual or spoken outputs (Bu et al., 2024). Compared to traditional pipeline architectures that sequentially chain Automatic Speech Recognition (ASR), Large Language Models (LLMs), and Text-To-Speech (TTS) components, end-to-end LSLMs offer significant advantages, including reduced latency, inherent error resilience, and more expressive speech synthesis capabilities (Ji et al., 2024).

Recent studies in LSLMs have focused on aligning speech modalities with text space through speech tokenizers (Zhang et al., 2023) or encoder-based approaches (Fang et al., 2024; Zhao et al., 2024; Chu et al., 2023). These cross-modal alignment strategies aim to harness the linguistic capabilities of pretrained LLMs while integrating speech processing functionalities (Cui et al., 2025).

However, significant performance disparities persist between LSLMs and conventional pipeline models in semantic understanding tasks. Benchmark results from VoiceBench (Chen et al., 2024) reveal a striking contrast: the *Whisper-v3-large* + *LLaMA-3.1-8B* pipeline achieves 79.06 overall score, while its LSLM counterpart *LLaMA-Omni* (Fang et al., 2024) scores merely 37.51. This pattern continues in Uro-bench (Yan et al., 2025) evaluations, where the *Whisper-v3-large* + *Qwen2-7B-Instruct* pipeline attains 78.13 overall score compared to *Freeze-Omni*'s (Wang et al., 2024) 48.28, despite both systems employing the same underlying LLM. Notably, while Uro-bench's dependence on transcribed speech outputs might inherently favor pipeline architectures, the magnitude of these performance drops remains substantial and warrants investigation.

Nevertheless, contemporary investigations into LSLMs remain largely confined to engineering practices, adopting unverified integrated solutions spanning training stages, dataset scales, parameter-efficient strategies, and multimodal objectives without systematic analysis of their individual contributions or synergistic effects (Chu et al., 2024; Zhong et al., 2024; Liu et al., 2025b). This practice of design-by-intuition raises critical concerns, as in-

validated architectural choices may inadvertently exacerbate the modality alignment discrepancy.

In this work, we take the persistent performance gap between end-to-end LSLMs and traditional ASR+LLM pipeline systems as our starting point. We systematically reproduce and quantify this discrepancy across various LLM backbones and training strategies, and, for the first time, empirically reveal the underlying mechanisms behind the performance difference. Specifically, after speech-text alignment training, a clear and consistent performance gap exists between text and speech inputs within the same model.

To gain insight into the modality gap, we systematically analyze the similarity between speech and text representations at both sequence and token levels, aiming to reveal the mechanisms of speech-text alignment within LSLMs. At the sequence (coarse-grained) level, we observe that as representations propagate through deeper layers of the model, their cosine similarity increases steadily, reflecting progressive directional alignment. In parallel, the Euclidean distance between modalities also increases, indicating a divergence in magnitude that likely reflects modality-specific characteristics learned by the model. At the token (fine-grained) level, we find that the model develops a spontaneous monotonic alignment pattern between speech and text tokens, indicating consistent local correspondence across modalities.

In addition, our study systematically examines the relationship between internal representation similarity and the modality gap exhibited on evaluation benchmarks. A clear linear correlation is observed at both the sequence and token levels, suggesting that the nature of internal cross-modal alignment is closely related to the performance disparity between speech and text inputs.

These observations are further examined through targeted intervention experiments, where speech token embeddings along the alignment path are modified using either angle projection or length normalization. We find that such interventions can improve performance on challenging cases from the `sd-qa` subset of VoiceBench.

Our contributions are threefold: (1) We systematically identify that the primary source of the performance gap between LSLMs and pipeline systems lies in the modality difference between speech and text inputs. (2) We analyze internal representations and find that the modality gap is closely linked to the similarity between speech and text representations at both sequence and token levels. (3) We provide the first empirical evidence that targeted interventions on speech representations can improve speech input performance on challenging cases.

By focusing on understanding and revealing the mechanisms behind modality alignment, our work offers a deeper exploration of the factors influencing LSLM performance. This approach not only addresses the current performance discrepancy but also paves the way for future advancements in integrating speech modalities into LLMs.

2 Related Work

2.1 Speech-Text Alignment in LSLMs

The alignment between speech and textual modalities is crucial for the performance of LSLMs. Recent studies have explored five distinct methodologies for this task (Ji et al., 2024). The latent space mapping approach, exemplified by *Qwen2-Audio* (Chu et al., 2024), *SALMONN* (Tang et al., 2023), and *VITA* (Fu et al., 2024), uses a joint audio encoder-adapter architecture to directly project speech inputs into the LLM's latent textual space. This paradigm effectively reduces computational overhead by compressing the audio sequence length via the audio adaptor module. Meanwhile, it also preserves the LLM's inherent reasoning capabilities and has demonstrated competitive performance across multiple benchmarks.

Additionally, *SpeechGPT* (Zhang et al., 2023) adopts modality chaining by discretizing speech into symbolic units and expanding the LLM's vocabulary, while *GLM-4-Voice* (Zeng et al., 2024) and *Moshi* (Défossez et al., 2024) utilize interleaved text-speech tokens and parallel generation architectures, respectively. *SyncLLM* (Veluri et al., 2024), *IntrinsicVoice* (Zhang et al., 2024b), *Align-SLM* (Lin et al., 2024), and *OmniFlatten* (Zhang et al., 2024a) pioneers direct speech-to-speech interaction without textual intermediates. Although significant progress has been made with these methodologies in existing research, their performance on audio processing tasks remains suboptimal.

2.2 Modality Gap Phenomenon

The modality gap phenomenon was systematically analyzed in (Liang et al., 2022), which demonstrated that this issue persists across a wide range of multimodal models. The gap is primarily attributed to the cone effect, where embeddings from different modalities occupy distinct subspaces, leading

to misalignment and degraded cross-modal performance.

In the context of speech translation, the presence and impact of the modality gap is investigated by (Fang and Feng, 2023) and (Han et al., 2023). Han et al. observed that the modality gap emerges during the early stages of fine-tuning. Fang et al. finding that modality misalignment can lead to divergent predictions and performance degradation compared to text-only machine translation systems. They quantified the gap using cosine similarity between speech and text embeddings, offering empirical evidence of substantial representational divergence.

2.3 Analysis of Multimodal Representations

Previous works have explored a variety of methods to accurately quantify the similarity between multimodal representations. Canonical Correlation Analysis (CCA) and its deep learning variants, such as SVCCA and PWCCA (Raghu et al., 2017; Morcos et al., 2018), have been widely used to capture linear and non-linear relationships of representations across different modalities.

To address the challenges of comparing high-dimensional representations, Centered Kernel Alignment (CKA), a robust similarity measure to identify correspondences between neural network representations, was proposed (Kornblith et al., 2019). In subsequent work, it has been further demonstrated that a simple, sample-wise cosine similarity metric can effectively capture layer-wise similarity in transformer models, closely aligning with the results of CKA while offering computational efficiency (Jiang et al., 2024).

Additionally, the Wasserstein distance between paired speech and text embeddings has also been used to measure cross-modal consistency (Liu et al., 2025a). Furthermore, the Gramian Representation Alignment Measure (GRAM) is designed to evaluate the alignment of multiple modalities simultaneously (Cicchetti et al., 2024). Both methods have been integrated into model training and have proven effective in enhancing the alignment between representations of different modalities.

3 Preliminary

This section investigates the performance degradation of LSLMs in processing speech inputs compared to their base models' performance on text inputs. Through comprehensive experiments conducted on multiple LLM backbones using both full-parameter and LoRA fine-tuning methods, we find that the primary contributor to this modality gap is the suboptimal alignment between textual and auditory modalities in LSLMs.

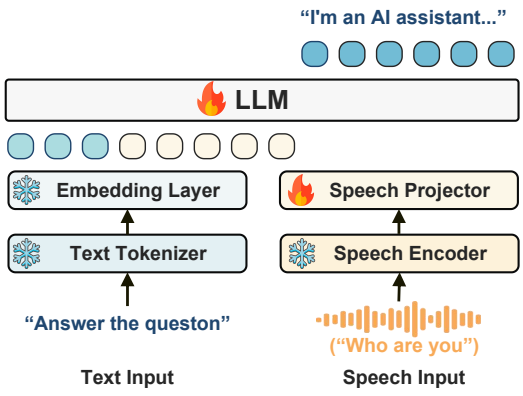

Figure 1: The Model Architecture.

3.1 Model Architecture

As illustrated in Figure 1, the architecture of our LSLMs comprises three core components: a speech encoder, a speech adaptor, and an LLM backbone. The speech signal is first encoded by the speech encoder into a latent representation. Subsequently, the speech adaptor compresses the temporal length of the speech sequence by a factor of N to reduce computational overhead. For text inputs, token embeddings are generated through the LLM backbone's embedding layer. The text token embeddings and compressed speech embeddings are then concatenated and fed into the LLM backbone to produce the final textual response.

In our experiments, *Whisper-large-v3*, a widely used ASR model, serves as the speech encoder. The speech adaptor is implemented as a lightweight module with two fully connected (FC) layers. We conducted experiments with various LLM backbones, including *LLaMA-3.2-3B-Instruct*, *LLaMA-3.1-8B*, *Qwen2.5-1.5B-Instruct*, and *Qwen2.5-7B-Instruct*. For brevity, henceforth we omit the suffix "Instruct" when referring to these model variants.

3.2 Experiment Setups

Our training dataset is constructed following the framework of *Ke-Speech-Chat* (Zhao et al., 2024), exclusively comprising single-turn dialogue samples. Each sample includes both speech and text instructions, as well as a text response. We refined the raw text using *Qwen2.5-72B-Instruct* (Yang et al., 2024; Team, 2024), aligning it with natural

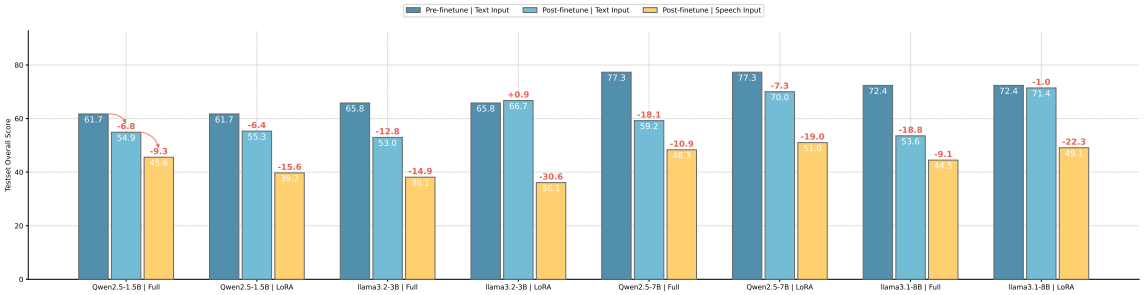

Figure 2: Overall Score (%, ↑) on the test set of 8 LSLMs at epoch 2 under different training methods and modality inputs. The red numbers on each bar show the score difference relative to the bar immediately to its left.

conversational patterns observed in real-world scenarios. Subsequently, a three-stage filtering mechanism was applied to purify the data, targeting safety, semantic clarity, and linguistic naturalness. The speech instruction-response pairs were synthesized using *CosyVoice* (Du et al., 2024). Based on automated transcription via *Whisper-large-v3* (Radford et al., 2022), speech samples exceeding a Word Error Rate (WER) threshold of 0.1 are discarded. Finally, Our training dataset contains 637,283 samples, with speech instructions totaling 1,604 hours.

All LSLMs are trained for 2 epochs on our training dataset using the AdamW optimizer with a peak learning rate of 2e-5. For LoRA, we set $r = 8$, $\alpha = 4.0$, and the dropout rate to 0.1. All experiments were conducted on a distributed setup with 2 nodes, each equipped with 8 NVIDIA A100 GPUs. Training a single model requires approximately 384 GPU hours on this setup.

For evaluation, we adopt five subsets of the *VoiceBench* dataset (Chen et al., 2024)—`AdvBench`, `IFEval`, `OBQA`, `MMSU`, and `sd-qa`—yielding a total of 4,947 test samples after filtering. These subsets collectively cover 93% of the full VoiceBench and are particularly suitable for robust evaluation as their metrics do not require additional LLMs, thereby minimizing variability. All evaluations strictly adhere to the official VoiceBench evaluation protocol to ensure consistency and reproducibility.

3.3 Results and Analysis

We evaluate the performance of LSLMs after full-parameter and LoRA fine-tuning on the 4 models introduced in Section 3.1, each tested under both speech and text modalities. Detailed results are provided in Appendix A.1. Figure 2 presents the performance of each model at epoch 2 on both text and speech inputs, alongside the corresponding base model's text-only performance.

For the data shown in Figure 2, on average, LSLMs exhibit a 25% performance decline on speech inputs relative to their base models on text. This decline can be attributed to two factors: (1) fine-tuning–induced erosion of reasoning and generation capabilities, with an average drop of 8.79%; and (2) suboptimal speech–text alignment, with an average drop of 16.46%. Given the variety of model sizes and tuning methods evaluated, this trend appears general. This phenomenon indicates that the observed performance degradation stems primarily from the speech–text modality gap, and that bridging this gap is crucial to enhance LSLM speech processing.

4 Empirical Analysis of Coarse-grained Speech-Text Representations

In this section, we examine the dynamic relationship between text and speech modality representations at a coarse-grained sequence level using similarity measurement techniques. Our analysis uncovers consistent patterns across various LLM architectures and training paradigms. Through extensive experimentation, we observe a strong linear correlation between representation alignment and performance disparities across modalities, particularly under LoRA fine-tuning, highlighting the predictive value of embedding similarity for modality gap estimation.

4.1 Methodology

Given a set of N speech-text query pairs $\{(x_i^s, x_i^t)\}_{i=1}^N$, where x_i^s denotes the speech input and x_i^t its corresponding text transcription, we process each sample through the model as illustrated in Figure 3.

For the speech modality, the input x^s is encoded by the speech encoder and linear projector, resulting in an initial embedding sequence $h_0^s \in \mathbb{R}^{T_s \times d}$,

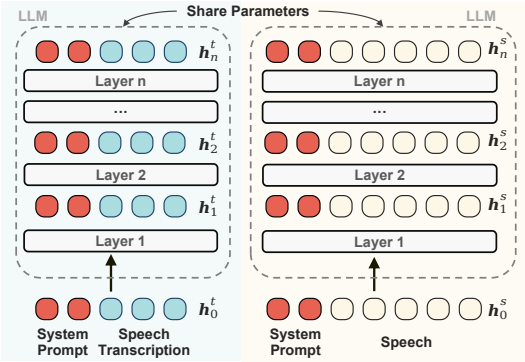

Figure 3: Overview of coarse-grained speech-text representations computation.

where T_s is the number of speech frames and d is the hidden dimension. This sequence, along with a system prompt, is fed into an L-layer model, yielding layer-wise representations $h_l^s \in \mathbb{R}^{T_s \times d}$, where $l \in \{1, \ldots, L\}$ indexes the model layer. Similarly, for the text modality, the input x^t is tokenized and embedded, producing $h_0^t \in \mathbb{R}^{T_t \times d}$, where T_t is the length of the token sequence. The corresponding layer-wise representations are $h_l^t \in \mathbb{R}^{T_t \times d}$.

To quantify the relationship between speech and text representations, we employ two similarity metrics, denoted in a unified manner as $f^{(\cdot)}(x, y)$, where $(\cdot)$ indicates the choice of metric (cos: cosine similarity, d: Euclidean distance):

$$f^{(\cos)}(x, y) = \frac{x^\top y}{\|x\|\|y\|}, \qquad f^{(\mathrm{d})}(x, y) = \|x - y\|_2.$$

For each layer l, we first compute the mean representation over the sequence dimension for both modalities:

$$\bar{h}_l^s = \frac{1}{T_s} \sum_{i=1}^{T_s} h_{l,i}^s, \qquad \bar{h}_l^t = \frac{1}{T_t} \sum_{j=1}^{T_t} h_{l,j}^t,$$

where $h_{l,i}^s$ and $h_{l,j}^t$ denote the i-th and j-th frame or token embedding at layer l for the speech and text modalities, respectively. The global relationship between modalities at each layer is then assessed by computing $f_l^{(\cos)}\left(\bar{h}_l^s, \bar{h}_l^t\right)$ and $f_l^{(\mathrm{d})}\left(\bar{h}_l^s, \bar{h}_l^t\right)$.

4.2 Sequence-level Speech-Text Representation Dynamics

We evaluate the sequence-level similarity at each layer for all test samples using the methodology outlined in Section 4.1. Figure 4 summarizes these results for various LSLMs under different training regimes, with cosine similarity shown in blue and Euclidean distance in orange. Each subplot corresponds to a specific model configuration, and each curve within a subplot represents a distinct training checkpoint, depicting the layer-wise similarity metrics.

Several notable patterns are observed in the similarity dynamics. For **cosine similarity**, all models demonstrate a consistent increase as the network depth grows, indicating progressively stronger alignment between speech and text representations in deeper layers. Moreover, later training checkpoints consistently yield higher similarity scores across all layers, reflecting improved cross-modal alignment as training advances.

In contrast, **Euclidean distance** gradually increases in the shallow layers and accelerates in the deeper layers. While models with more extensive training tend to exhibit slightly lower Euclidean distances overall, the upward trend with increasing depth remains consistent, suggesting growing representational magnitude divergence despite directional convergence.

These trends suggest that deeper layers and extended training foster improved alignment in representational direction (cosine similarity), while preserving modality-specific distinctions in magnitude (Euclidean distance). This alignment pattern could facilitate effective multimodal integration while preserving essential characteristics of each modality.

4.3 Correlation Between Representation Similarity and Modality Gap

To analyze the relationship between representation similarity and downstream performance, we compute a scalar similarity score for each model by averaging similarity across layers:

$$\bar{f}^{(\cdot)} = \frac{1}{L} \sum_{l=1}^{L} f_l^{(\cdot)}\left(\bar{h}_l^s, \bar{h}_l^t\right),$$

where $(\cdot)$ denotes either cosine similarity or Euclidean distance.

We quantify the modality gap as the drop in benchmark scores between text and speech inputs, as:

$$GAP = M^t - M^s,$$

where M^t and M^s are overall benchmark scores obtained from text and speech inputs, respectively.

Figure 5 shows the linear relationship between similarity and GAP. Each point corresponds to a

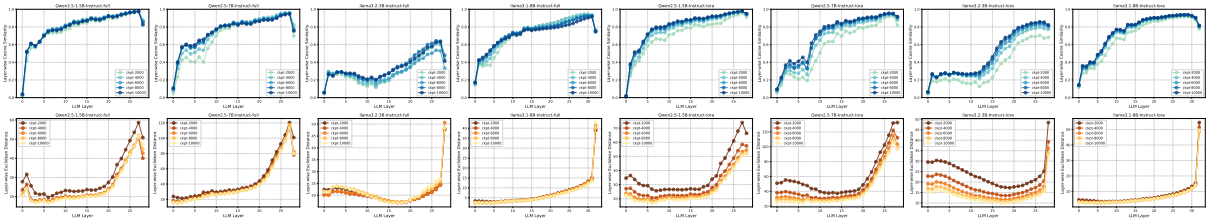

Figure 4: Cosine similarity (blue) and Euclidean distance (orange) between speech and text representations across model layers and checkpoints.

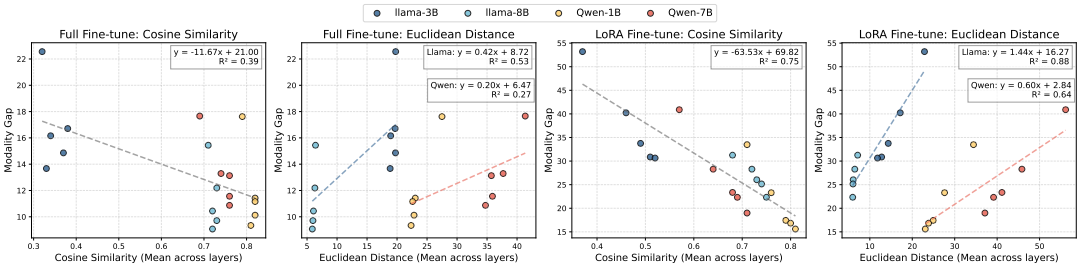

Figure 5: Linear relationship between cross-modal similarity and modality performance gap (GAP).

model checkpoint, with the R^2 value indicating the strength of correlation.

Key Findings. Under LoRA fine-tuning, a strong linear relationship is observed between cosine similarity and GAP ($R^2 = 0.75$), suggesting that better cross-modal alignment leads to smaller performance disparities. Although Euclidean distance shows a weaker correlation overall, it becomes more pronounced within specific model families ($R^2 = 0.64$ for *Qwen* and $R^2 = 0.88$ for *Llama*).

For full-parameter fine-tuning, the same trend persists but with reduced strength: $R^2 = 0.39$ for cosine similarity, and $R^2 = 0.53$ and 0.27 for Euclidean distance in *Llama* and *Qwen*, respectively.

Analysis. These findings empirically validate the connection between internal cross-modal representations and performance-level modality gaps. The stronger correlations observed in LoRA-tuned models may stem from the constrained low-rank adaptation, which preserves the integrity of pretrained text representations while facilitating targeted speech-text alignment. In contrast, full fine-tuning grants more representational flexibility, potentially introducing overfitting that weakens this correlation. Consequently, representation similarity serves as a more reliable predictor of modality performance under LoRA than under full-parameter.

5 Empirical Analysis of Finer-grained Speech-Text Representations

Building on the coarse-grained sequence-level analysis in the previous section, this section focuses on token-level alignment patterns, examining **the role and contribution of each token in modality alignment**. We will begin with case studies, and subsequently introduce more detailed quantitative metrics to facilitate a finer-grained investigation. Through correlation analysis and intervention experiments, we explore the relationship between token-level alignment and downstream task performance, thereby further revealing the underlying speech-text alignment mechanisms in LSLMs.

5.1 Monotonic Patterns in Token-wise Similarity Matrices

Observations At each layer l, we construct a token-wise similarity matrix $A_l^{(\cdot)} \in \mathbb{R}^{T_s \times T_t}$, defined as:

$$[A_l^{(\cdot)}]_{i,j} = f^{(\cdot)}(h_{l,i}^s, h_{l,j}^t),$$

where $f^{(\cdot)}$ denotes the selected similarity metric. Across all models and training paradigms, we consistently observed that **the token-wise similarity matrix typically exhibits extreme values along a nearly monotonic path**. As shown in Fig. 6, with the increase in text token index, there is a monotonic alignment path in the speech frame sequence along which the similarity (or distance) values are

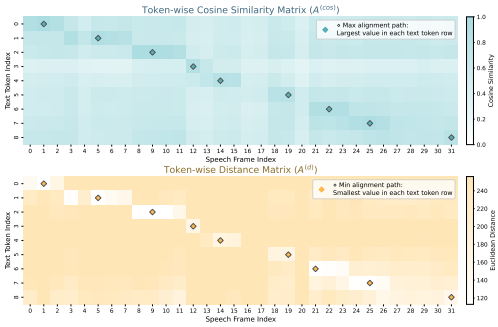

Figure 6: Token-wise similarity and distance matrices for a representative layer, with diamond markers indicating the alignment path.

locally maximized (or minimized). This monotonic path does not strictly align with the main diagonal, but reflects the actual temporal alignment structure between speech and text modalities.

Statistical Quantification To systematically quantify this alignment pattern, for each text token j, we identify the index of the speech frame with maximal similarity or minimal distance as:

$$i_j^* = \begin{cases} \arg\max_i [A_l^{(\cos)}]_{i,j} & \text{if } f^{(\cdot)} = f^{(\cos)} \\ \arg\min_i [A_l^{(\mathrm{d})}]_{i,j} & \text{if } f^{(\cdot)} = f^{(\mathrm{d})} \end{cases}$$

This process produces an alignment path between the text and speech sequences. To verify the presence of monotonicity in these alignments, we use the Spearman rank correlation coefficient between text token indices and their aligned speech frame indices as the evaluation metric. Detailed statistics are provided in Appendix A.2. At the final training epoch across all models, the average Spearman coefficient is 0.85 for cosine similarity and 0.70 for Euclidean distance. The proportion of tokens with perfectly identical alignment paths under both similarity measures is 0.59, indicating substantial consistency in the alignment results.

Mechanism Analysis The widespread emergence of this monotonic pattern suggests that the model not only aligns modalities globally, but also **spontaneously learns a soft, monotonic alignment between speech frames and text tokens at the token level**. Importantly, this alignment pattern emerges automatically in end-to-end speech-text alignment tasks, reflecting the model's ability to capture and map the temporal structure of speech to the semantic structure of text in a robust manner.

5.2 Alignment Path Score

Based on the observed token-level alignment patterns, we propose the **Alignment Path Score** (APS) to quantify the strength of speech-text alignment at the token level. Specifically, APS is defined as:

$$\mathrm{APS}^{(\cdot)} = \frac{1}{L \times T_t} \sum_{l=1}^{L} \sum_{j=1}^{T_t} [A_l^{(\cdot)}]_{i_j^*,j},$$

where L denotes the number of layers, T_t is the number of text tokens, and $[A_l^{(\cdot)}]_{i_j^*,j}$ represents the maximal similarity (or minimal distance) along the alignment path for each token.

We systematically evaluate the relationship between APS and GAP defined in Section 4.3 on the LSLMs using a linear regression analysis. As shown in Figure 7, LoRA-trained LSLMs yield higher R^2 values for both cosine (0.81 vs. 0.75) and Euclidean APS (0.72 vs. 0.64 for *Qwen*; 0.95 vs. 0.88 for *Llama*) compared to previous baselines, indicating stronger linear correlations with GAP. Under full-parameter finetuning, the correlation between APS and GAP is similar to that of sequence-level metrics, with both showing low R^2 values. As previously suggested, the greater training noise and instability in this setting may limit the explanatory power of both sequence-level and token-level alignment metrics.

These results suggest that APS offers a more direct and sensitive measurement of the relationship between alignment quality and downstream performance. The stronger correlation between APS and GAP highlights that fine-grained, token-level alignment is the key mechanism underlying LSLM speech understanding.

5.3 Intervention Experiments: Probing the Causal Role of Token-level Alignment

As demonstrated in the previous section, our analyses revealed a strong correlation between the token-level alignment score (APS) and the modality gap in model performance (GAP). However, correlation does not necessarily imply causation. To further investigate whether the token-level alignment mechanism causally affects the speech understanding ability of LSLMs, we conducted a series of targeted intervention experiments.

Specifically, we focused on the `sd-qa` subset of VoiceBench and selected both *Qwen2.5-7B* and *Llama3.1-8B* models, each under LoRA and full-parameter fine-tuning settings. For each sample,

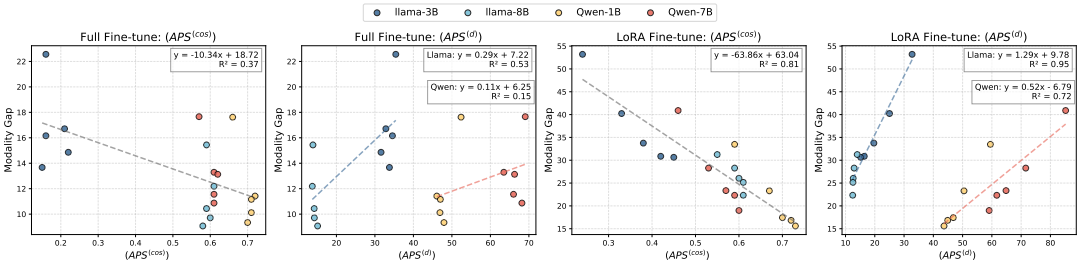

Figure 7: Relationship between the proposed Alignment Path Score and modality performance gap (GAP).

Table 1: Performance (%, ↑) on the `sd-qa` subset under different token-level intervention strategies. "*Bottom3*": Only the three least-aligned tokens are modified; "*All*": The entire alignment path is modified. For each row, results that outperform the corresponding Original are typeset in bold.

Model	Strategy	Original	Angle projection		Length normalization	
			Bottom3	*All*	*Bottom3*	*All*
Qwen2.5-7B	Full	33.45	**38.52**	**38.70**	**37.25**	32.73
	LoRA	38.88	**40.51**	38.88	38.16	31.10
Llama3.1-8B	Full	36.53	32.91	34.00	31.65	30.74
	LoRA	45.75	**49.19**	**47.74**	42.86	39.96

we first used the APS path to identify the three speech tokens with the lowest alignment scores (*bottom3*) as well as all tokens along the alignment path (*All*). We then applied two types of interventions: (1) **Angle projection**, where the selected speech token embeddings were projected to have the same direction as their corresponding text token embeddings; and (2) **Length normalization**, where the norm of the speech token embeddings was scaled to match that of the corresponding text tokens. We evaluated the downstream QA accuracy before and after intervention.

As shown in Table 1, Angle Projection yields improvements or maintains performance in 6 out of 8 intervention settings, demonstrating that increasing the angular similarity of token-level text and speech representations can enhance downstream outcomes. For LoRA fine-tuned models, applying angle projection to either the *Bottom3* or *All* alignment-path tokens consistently improves results. Notably, intervening on only the *Bottom3* tokens leads to more robust gains, with *Llama3.1-8B* improving by 7.52% and *Qwen2.5-7B* by 4.19%. In contrast, length normalization provides improvement in only one case, with performance declining in the remaining settings, indicating an overall detrimental effect on LSLM's speech sequence modeling.

Further case analysis reveals that angular or length-based interventions on speech tokens can correct instances where the model previously failed to generate accurate responses for speech input while succeeding for the corresponding text input. These corrections fall into two primary categories: (1) resolving semantic misunderstandings arising from misinterpretation of spoken content, and (2) rectifying factual errors despite correct semantic parsing. Representative examples illustrating both types of correction are provided in Appendix A.3, underscoring the potential of token-level interventions to enhance both linguistic comprehension and factual consistency for spoken queries.

6 Conclusion

This work systematically investigates the modality gap in LSLMs, defined as the performance disparity between speech and text inputs within the same trained model. To uncover the mechanisms behind this gap, we analyze speech-text alignment at both sequence and token levels. Sequence-level analysis tracks representation similarity across layers and training, establishing its linear relationship with the modality gap. At the token level, we reveal word-frame alignment structures and propose the Alignment Path Score, which shows a stronger correlation with the proposed modality gap. Targeted intervention experiments further demonstrate that improving token-level alignment can enhance speech inference accuracy. This study deepens understanding of how large language models process and comprehend spoken language.

Limitations

While this study offers a systematic empirical analysis of the modality gap and alignment mechanisms in large speech language models, several limitations should be acknowledged. Our findings are based on experiments with specific model architectures and alignment frameworks, so their generalizability to other speech encoders, alignment paradigms, or larger-scale models remains to be further validated. Moreover, the evaluation primarily focuses on English single-turn dialogue tasks with synthetic speech, potentially limiting the applicability of our conclusions to more diverse languages, conversational settings, or real-world noisy speech inputs. Finally, the proposed intervention strategies are applied post hoc during inference rather than integrated into model training; as such, we view them as analytical tools to deepen understanding of modality alignment, and future work may explore how these insights can inform the development of more robust and generalizable training strategies.

References

Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. 2024. Roadmap towards superhuman speech understanding using large language models. *Preprint*, arXiv:2410.13268.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

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

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. 2024. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. *Preprint*, arXiv:2410.03751.

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

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. *Preprint*, arXiv:2305.08706.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, and 1 others. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Yuchen Han, Chen Xu, Tong Xiao, and Jingbo Zhu. 2023. Modality adaption or regularization? a case study on end-to-end speech translation. *arXiv preprint arXiv:2306.07650*.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. 2024. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Preprint*, arXiv:2203.02053.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. 2024. Align-slm: Textless spoken language models with reinforcement learning from ai feedback. *arXiv preprint arXiv:2411.01834*.

Henglyu Liu, Andong Chen, Kehai Chen, Xuefeng Bai, Meizhi Zhong, Yuan Qiu, and Min Zhang. 2025a. Adaptive inner speech-text alignment for llm-based speech translation. *Preprint*, arXiv:2503.10211.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025b. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *Preprint*, arXiv:2502.04328.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents. *arXiv preprint arXiv:2409.15594*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *Preprint*, arXiv:2411.00774.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: A comprehensive benchmark for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chaohong Tan, Zhihao Du, and 1 others. 2024a. Omniflatten: An end-to-end gpt model for seamless voice conversation. *arXiv preprint arXiv:2410.17799*.

Xin Zhang, Xiang Lyu, Zhihao Du, Qian Chen, Dong Zhang, Hangrui Hu, Chaohong Tan, Tianyu Zhao, Yuxuan Wang, Bin Zhang, and 1 others. 2024b. Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities. *arXiv preprint arXiv:2410.08035*.

Shuaijiang Zhao, Tingwei Guo, Bajian Xiang, Tongtang Wan, Qiang Niu, Wei Zou, and Xiangang Li. 2024. Advancing speech language models by scaling supervised fine-tuning with over 60,000 hours of synthetic speech dialogue data. *arXiv preprint arXiv:2412.01078*.

Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, Shaozuo Yu, Sitong Wu, Eric Lo, Shu Liu, and Jiaya Jia. 2024. Lyra: An efficient and speech-centric framework for omni-cognition. *Preprint*, arXiv:2412.09501.

A Appendix

A.1 Detailed Comparison of Model Performance with Speech and Text Inputs across Training Strategies

Table 2 presents the evaluation results of our models under different training paradigms and checkpoints. For each model and training strategy, we report the performance on both speech input and text input across multiple benchmark subsets, as well as their respective overall scores. Additionally, we provide the GAP metric, defined as the difference between the overall text input and speech input performance. This comprehensive comparison allows us to assess the alignment and robustness of various models and training approaches with respect to both input modalities.

A.2 Analysis of Alignment Path Monotonicity and Consistency

This section details the measurement of alignment path statistics, as introduced in Section 5.1. We report these statistics across different training stages, model scales, and training strategies. As shown in Figure 8, we consider three metrics: (1) alignment path monotonicity based on the cosine similarity matrix, which reflects the degree of order in the alignment between text tokens and speech frames; (2) alignment path monotonicity based on the Euclidean distance matrix, defined in a similar manner but using Euclidean distances for alignment construction; and (3) token-level alignment path consistency, defined as the proportion of tokens whose aligned speech frame indices are identical under both similarity measures.

Model	Strategy	Param	Steps	Speech Input (%, ↑)						Text Input (%, ↑)						*GAP* (↓)
				AdvBench	IfEval	OBQA	MMSU	sd-qa	Overall	AdvBench	IfEval	OBQA	MMSU	sd-qa	Overall	
Qwen2.5-1.5B	Full	1.5B	2,000	77.50	15.55	25.27	26.25	31.83	35.28	96.73	20.16	61.10	42.06	44.44	52.90	17.62
			4,000	95.77	14.01	44.40	29.99	31.10	43.06	98.85	20.64	69.01	44.99	38.89	54.48	11.42
			6,000	94.04	12.39	45.93	32.60	29.84	42.96	97.69	18.97	69.45	45.02	39.47	54.12	11.16
			8,000	95.77	13.67	43.52	31.36	30.02	42.87	98.65	18.44	68.13	45.25	34.45	52.99	10.12
			10,000	98.46	13.36	49.45	33.12	33.45	45.57	99.04	18.64	69.89	46.39	40.60	54.91	9.34
	LoRA	9M	2,000	32.17	11.40	25.71	26.61	23.33	23.84	88.08	32.16	73.85	52.60	39.96	57.33	33.48
			4,000	59.35	13.67	30.99	26.28	27.85	31.63	84.42	24.27	74.51	74.51	39.42	54.93	23.30
			6,000	80.00	14.22	36.04	27.00	29.48	37.35	88.08	21.90	74.07	51.89	37.97	54.78	17.43
			8,000	83.08	13.81	36.48	27.94	30.20	38.30	88.46	21.82	74.95	51.76	38.70	55.14	16.83
			10,000	84.42	13.46	40.00	30.32	30.38	39.72	88.08	21.56	75.38	51.43	40.14	55.32	15.60
Qwen2.5-7B	Full	7B	2,000	92.31	14.85	43.30	31.10	36.17	43.54	98.27	23.17	80.00	60.44	44.12	61.20	17.66
			4,000	96.54	16.14	54.29	33.96	33.82	46.95	99.62	23.29	78.68	58.36	41.23	60.23	13.29
			6,000	94.23	18.60	57.80	36.92	36.35	48.78	99.23	26.55	79.56	59.17	45.03	61.91	13.13
			8,000	97.31	16.58	58.68	35.88	32.19	48.13	99.04	23.80	75.16	58.65	41.77	59.69	11.56
			10,000	98.85	15.84	58.02	35.56	33.45	48.34	100.0	24.39	71.87	56.25	43.58	59.22	10.87
	LoRA	20M	2,000	67.12	14.16	27.91	26.28	35.44	34.18	99.04	62.30	87.47	69.06	57.50	75.08	40.89
			4,000	92.12	16.92	42.64	30.03	35.62	43.47	99.23	55.38	88.13	68.41	47.56	71.74	28.28
			6,000	94.81	22.10	48.13	34.16	37.07	47.25	99.04	50.79	87.25	67.86	48.10	70.61	23.35
			8,000	95.19	22.63	49.45	35.56	38.52	48.27	99.04	49.71	88.13	67.79	48.28	70.59	22.32
			10,000	96.92	25.42	56.26	37.67	38.88	51.03	99.04	48.19	87.69	67.24	47.92	70.02	18.99
Llama3.2-3B	Full	3B	2,000	75.38	12.14	23.30	25.08	27.12	32.61	98.27	26.13	63.52	45.22	42.68	55.16	22.56
			4,000	95.96	12.06	23.52	25.02	31.46	37.60	98.46	21.90	63.96	43.82	40.69	53.76	16.16
			6,000	99.04	12.97	27.47	25.47	34.90	39.97	99.23	22.43	61.32	42.55	42.68	53.64	13.67
			8,000	93.85	13.72	24.84	24.98	29.84	37.44	98.65	24.17	62.20	42.88	42.86	54.15	16.71
			10,000	98.65	11.60	22.64	24.98	32.73	38.12	99.23	23.88	59.12	40.53	42.13	52.98	14.86
	LoRA	12M	2,000	11.15	14.66	21.32	25.54	19.89	18.51	97.50	63.37	75.60	56.64	65.46	71.71	53.20
			4,000	58.08	12.18	22.42	23.19	35.62	30.30	98.46	62.64	76.92	55.50	59.13	70.53	40.23
			6,000	74.42	12.97	20.88	24.79	38.70	34.35	98.08	59.32	75.60	54.52	52.98	68.10	33.75
			8,000	84.42	13.52	19.12	25.37	38.52	36.19	98.08	55.39	77.14	54.39	50.27	67.05	30.86
			10,000	79.66	13.28	22.64	24.53	40.33	36.09	97.69	53.25	76.92	54.26	51.54	66.73	30.64
Llama3.1-8B	Full	8B	2,000	93.27	13.67	30.77	25.80	34.90	39.68	99.62	19.55	68.13	44.37	43.94	55.12	15.44
			4,000	99.23	14.23	31.43	27.03	36.89	41.76	99.62	18.11	67.91	41.09	43.04	53.95	12.19
			6,000	97.50	13.42	38.68	28.43	37.61	43.13	99.04	18.71	64.62	43.33	42.13	53.57	10.44
			8,000	98.85	12.70	33.63	27.26	37.97	42.08	99.81	17.03	58.90	39.66	43.58	51.79	9.71
			10,000	99.23	13.94	44.18	28.59	36.53	44.49	99.42	17.91	67.03	42.39	41.05	53.56	9.07
	LoRA	20M	2,000	85.96	16.60	27.69	27.62	43.76	40.33	90.38	67.68	81.10	64.09	54.61	71.57	31.25
			4,000	94.23	17.43	36.04	30.68	43.94	44.47	99.42	63.58	81.54	64.57	54.61	72.75	28.28
			6,000	93.46	19.16	41.10	31.10	45.57	46.08	99.42	61.10	81.10	64.18	54.79	72.12	26.04
			8,000	97.12	19.36	43.74	32.37	43.58	47.23	99.42	62.75	81.10	64.02	54.61	72.38	25.15
			10,000	96.92	21.12	49.01	32.69	45.75	49.10	99.42	59.16	81.10	64.31	53.16	71.43	22.33

Table 2: Comparison of Alignment Experiment Results: Speech and Text Input Performance Across Steps

Empirical results reveal the following trends: (1) Both alignment path monotonicity metrics exhibit an overall increasing tendency as training progresses, suggesting that the model incrementally acquires more structured and monotonic alignments. (2) The monotonicity measured via cosine similarity remains consistently higher than that based on Euclidean distance, indicating that cosine similarity may be more effective in capturing ordered relationships in high-dimensional spaces. (3) Token-level alignment path consistency also demonstrates a general upward trend during training, implying that the alignment paths derived from the two similarity measures become increasingly similar. These observations are consistent across different model scales and training strategies, underscoring the robustness and effectiveness of the learned alignment mechanism.

A.3 Case Analysis of Intervention Experiments

As shown in Table 3, Table 4, Table 5, and Table 6, we present representative cases from our intervention experiments in Section 5.3. In these tables, **angle** denotes interventions where the speech representation is projected towards the aligned text representation direction , while **length** refers to adjusting the representation norm to match that of the text. The **bot3** setting applies the intervention only to the three tokens along the alignment path with the lowest confidence, whereas **all** applies the intervention to all tokens along the alignment path.

We identify two main categories of cases to illustrate the effects of these interventions:

1. Semantic Misunderstanding from Spoken Input. In the first category, the model fundamentally misunderstands the meaning or entity referenced in the spoken input. For example, in Case 1 (Table 3), the model misinterprets "Brittany" as "Britain," leading to an irrelevant answer. Similarly, in Case 2 (Table 4), the spoken input is incorrectly recognized as a different entity, resulting in an erroneous response. After applying either angle or length-based interventions to key tokens along the alignment path, the model realigns its semantic representation with the ground-truth text, recovers the correct understanding, and generates accurate answers.

2. Factual Errors with Correct Semantic Parsing. In the second category, the model correctly

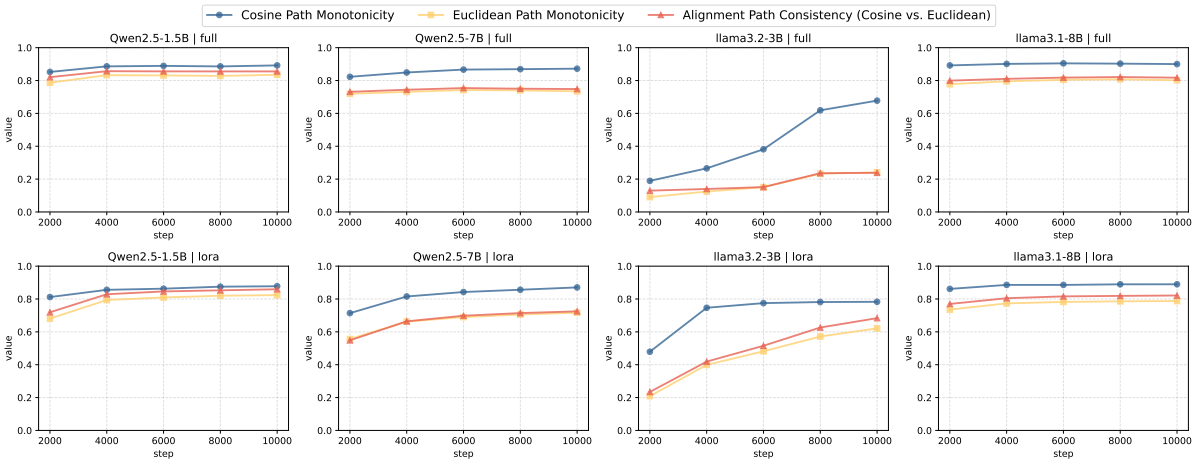

Figure 8: **Evolution of Alignment Path Monotonicity and Consistency.** For each model (*columns*) and training strategy (*rows*), we report the trajectory of three metrics: Cosine Path Monotonicity (blue), Euclidean Path Monotonicity (yellow), and Alignment Path Consistency (orange) across checkpoints during training. Higher monotonicity values indicate a stronger tendency toward monotonic alignments between text tokens and speech frames. Higher consistency values reflect greater agreement between alignment paths derived from the two similarity measures, suggesting stable and modality-agnostic alignment patterns.

parses the user's intent and the core question from spoken input, but produces factually incorrect or incomplete answers. For instance, in Case 3 (Table 5) and Case 4 (Table 6), the model demonstrates a correct understanding of the question topic, yet the factual details in its output are incorrect. Our interventions—by modifying the representation direction or norm along the alignment path—are also effective in these cases, guiding the model to generate factually correct responses consistent with the reference answers.

These case studies demonstrate that fine-grained interventions on the alignment path—either in embedding direction or norm—consistently improve the model's answer accuracy and robustness for spoken inputs. This effect is observed in both cases of semantic misinterpretation and factual knowledge errors, indicating that such interventions can enhance multi-modal alignment and enable more reliable knowledge retrieval from speech input.

Case 1
Question: What do they speak in Brittany?
Ref: Breton.
Text Input: In Brittany, the local language is Breton. It's a Celtic language that's still spoken by some people in the region. (✓) **Speech Input:** In Britain, the primary language spoken is English. However, there are also other languages spoken, especially in areas with diverse populations, such as Welsh in Wales, Scottish Gaelic in Scotland, and Irish in Northern Ireland. (✗) **Speech Input (Angle Bot3):** Breton is a Celtic language that's still spoken in parts of France, particularly in the region of Brittany. It has a unique sound and rhythm that's quite distinct from other languages. (✓) **Speech Input (Length Bot3):** in Scotland, they speak Scottish Gaelic. In Northern Ireland, they speak Northern Irish, which is a dialect of Irish. In Wales, they speak Welsh. And in Cornwall, they speak Cornish. Each place has its own unique language and cultural heritage. (✗) **Speech Input (Angle All):** In Brittany, they primarily speak French. However, there are some areas where you might also hear Breton, which is a Celtic language. But don't worry, most people there will be speaking French. (✓) **Speech Input (Length All):** Cockney, which is a type of East London accent. It's known for its unique way of pronouncing certain letters and sounds, often making it quite distinct from other accents in the United Kingdom. (✗)

Table 3: Case 1 from intervention experiments on the sd-qa dataset.

Case 2
Question: Who is Sean Hannity?
Ref: An American talk show host and conservative political commentator.
Text Input: Sean Hannity is an American conservative political commentator and radio host. He's best known for hosting his own show on Fox News, where he discusses politics, current events, and social issues. He's also a popular author and has written several books. (✓) **Speech Input:** Shawn Hanity is an American actor and comedian. He's best known for his roles in television shows like Saturday Night Live and Brooklyn Nine-Nine. He's also appeared in several films and has been involved in various other projects. (✗) **Speech Input (Angle Bot3):** Sean Hannity is a well-known American conservative political commentator and television personality. He hosts a popular radio show and has his own talk show on Fox News, where he often discusses current political issues and provides commentary on various topics. (✓) **Speech Input (Length Bot3):** Shan? He's a British comedian and actor known for his work in both stand-up and on television shows like The League of Gentlemen. (✗) **Speech Input (Angle All):** Sean Hannity is a well-known American conservative political commentator and television personality. He hosts a popular radio show and has his own talk show on Fox News, where he often discusses current political issues and supports Republican candidates. (✓) **Speech Input (Length All):** Sean Hannity is a well-known American conservative political commentator and television personality. He hosts a popular radio show and has his own talk show on Fox News, where he often discusses current political issues and provides commentary on various topics. Sean Hannity is also a best-selling author and has written several books. (✓)

Table 4: Case 2 from intervention experiments on the sd-qa dataset.

Case 3

Question: What's the biggest wasp?

Ref: Asian giant hornet.

Text Input: The largest wasp in the world is the Asian giant hornet. It can grow up to two and a half inches long, which is quite big compared to other wasps. (✓)
Speech Input: The largest wasp is the tarantula hawk wasp, which can grow up to two and a half inches long. These wasps are known for their impressive size and their ability to paralyze tarantulas to use as food for their larvae. (✗)
Speech Input (Angle Bot3): The largest wasp species is the giant hornet, also known as the Asian giant hornet. These impressive insects can have a wingspan of up to two inches and are found in various parts of Asia. They are known for their aggressive behavior and can be quite dangerous. (✓)
Speech Input (Length Bot3): It's called the Great Barrier Reef, and it's located off the coast of Australia. It's so big that it can even be seen from space! (✗)
Speech Input (Angle All): The largest wasp species is the golden paper wasp. They can build some of the largest and most complex nests among all wasp species, often using chewed wood fibers mixed with saliva to create a papery material. These nests can be quite impressive! (✗)
Speech Input (Length All): It's called the Great Barrier Reef, and it's located off the coast of Australia. It's so big that it can even be seen from space! (✗)

Table 5: Case 3 from intervention experiments on the sd-qa dataset.

Case 4

Question: Where does the story of American Horror Story: Murder House take place?

Ref: Los Angeles, California.

Text Input: The story of American Horror Story: Murder House takes place in Los Angeles, California. (✓)
Speech Input: The story of American horror story murder house takes place in the fictional town of Harmony, which is modeled after the real town of New Orleans. The house itself is located in the French Quarter, known for its rich history and eerie atmosphere. (✗)
Speech Input (Angle Bot3): American Horror Story Murder House is set in a fictional address in Los Angeles, California. The exact location isn't specified, but it's portrayed as being in the heart of the city. (✓)
Speech Input (Length Bot3): American Horror Story Murder House is set in a fictional address in Los Angeles, California. The exact location isn't specified, but it's portrayed as being in the heart of the city. (✓)
Speech Input (Angle All): American Horror Story Murder House is set in New York City. (✗)
Speech Input (Length All): American Horror Story Murder House is set in New York City. The story revolves around a haunted house in the heart of the city, where the main character, a young woman named Wendy, moves in with her husband and their son. The house has a dark history, and its past is gradually revealed throughout the series. (✗)

Table 6: Case 4 from intervention experiments on the sd-qa dataset.