Listen like a Teacher: Mitigating Whisper Hallucinations Using Adaptive Layer Attention and Knowledge Distillation

Kumud Tripathi*, Aditya Srinivas Menon*,
Aman Gaurav, Raj Prakash Gohil, Pankaj Wasnik
Sony Research India,
kumud.tripathi,aditya.menon,aman.gaurav,raj.gohil,pankaj.wasnik@sony.com

Abstract

The Whisper model, an open-source automatic speech recognition system, is widely adopted for its strong performance across multilingual and zero-shot settings. However, it frequently suffers from hallucination errors, especially under noisy acoustic conditions. Previous works to reduce hallucinations in Whisper-style ASR systems have primarily focused on audio preprocessing or post-processing of transcriptions to filter out erroneous content. However, modifications to the Whisper model itself remain largely unexplored to mitigate hallucinations directly. To address this challenge, we present a two-stage architecture that first enhances encoder robustness through Adaptive Layer Attention (ALA) and further suppresses hallucinations using a multi-objective knowledge distillation (KD) framework. In the first stage, ALA groups encoder layers into semantically coherent blocks via inter-layer correlation analysis. A learnable multi-head attention module then fuses these block representations, enabling the model to jointly exploit low- and high-level features for more robust encoding. In the second stage, our KD framework trains the student model on noisy audio to align its semantic and attention distributions with a teacher model processing clean inputs. Our experiments on noisy speech benchmarks show notable reductions in hallucinations and word error rates, while preserving performance on clean speech. Together, ALA and KD offer a principled strategy to improve Whisper's reliability under real-world noisy conditions.

Introduction

Recent progress in Automatic Speech Recognition (ASR) has been fueled by Transformer-based encoder-decoder architectures, such as Whisper, which effectively capture long-range dependencies and model complex acoustic and linguistic interactions (Yang et al. 2024; Tseng et al. 2024; Radford et al. 2023; Gulati et al. 2020). However, these models remain vulnerable to hallucinations: fluent yet semantically incorrect transcriptions. Hallucinations are particularly problematic in noisy or non-speech segments and can seriously undermine trust in speech systems, since they often escape detection by conventional metrics like the word error rate (WER) (Atwany et al. 2025; Frieske and Shi 2024). Recent research indicates that hallucinations often arise from misaligned internal representations in both the encoder and decoder when faced with noisy input conditions (Atwany et al. 2025). Many existing solutions focus on downstream techniques, such as pre-processing using voice activity detection (Bain et al. 2023), post-processing (Fang et al. 2022), and data augmentation (Barański et al. 2025), but they do not address the root cause in latent representations. To remedy this, we introduce a two-stage framework that first strengthens encoder robustness through Adaptive Layer Attention (ALA) and mitigates hallucinations using a multi-objective knowledge distillation (KD) strategy.

To enhance the robustness of encoders in noisy conditions, we introduce Adaptive Layer Attention (ALA), a dynamic fusion mechanism that utilises the hierarchical representations present within Transformer encoders. Since each transformer encoders comprise multiple layers, each capturing different levels of phonetic, lexical and semantic information (Barakat 2024). Therefore, relying solely on the final layer, as traditional ASR systems do, can result in the loss of intermediate features. ALA addresses this issue by performing inter-layer correlation analysis, which groups layers into semantically coherent blocks. It then applies a learnable multi-head attention (MHA) mechanism (Vaswani et al. 2017) to dynamically fuse these block representations during both training and inference time. This enables segment-wise selection of the most informative layer block, enhancing contextual modelling under noisy conditions. By adaptively fusing information across the encoder hierarchy, ALA produces richer and more stable representations, reducing dependence on any single layer and improving overall noise robustness.

To further reduce hallucinations beyond improvements at the encoder level, we introduce a Multi-Objective Knowledge Distillation (MOKD) strategy that supervises decoder behaviour under noisy conditions. While the ALA enhances the quality of input representations, hallucinations can still occur if the decoder misinterprets noisy segments. To address this, we fine-tune the ALA-augmented encoder-decoder model using a clean-speech trained ASR model as the teacher, which provides stable representations and attention distributions. During the training of the student model on noisy inputs, we use a composite loss function that optimizes three objectives: (1) cross-entropy loss for transcription accuracy, (2) cosine similarity loss between the teacher

*These authors contributed equally.

and student representations at the final encoder and decoder layers to promote semantic and contextual alignment, and (3) Mean Squared Error (MSE) loss on the decoder's cross-attention maps to transfer the teacher's attention behavior.

Our combined framework (MOKD) first equips the encoder with noise-aware adaptive attention, then supervises the decoder's behaviour to emulate clean speech patterns. This dual-stage design tackles hallucination at both the representation and attention levels. Related work in distillation for machine translation and speech recognition has shown that transferring attention alignment between models improves robustness and generalisation (Hentschel et al. 2024; Tseng et al. 2024; Yoon 2025; Nguyen et al. 2025). Building on these insights, our approach aligns both semantic representations and attention behaviours across teacher and student models. This results in improved WER and better semantic consistency, measured via metrics such as SeMaScore (Sasindran, Yelchuri, and Prabhakar 2024) on noisy speech datasets, while maintaining performance in clean speech. Empirical analysis demonstrates a substantial reduction in hallucination and more stable cross-attention patterns of the decoder. Our key contributions are:

1. We propose ALA over the Whisper encoder to enhance robustness under noisy speech conditions by incorporating low-level features from the intermediate layers.
2. We introduce a multi-loss KD strategy that combines cross-entropy, encoder/decoder cosine similarity, and decoder attention MSE to align intermediate representations with a clean teacher model.
3. We train the student model exclusively on noisy data while distilling knowledge from a clean-data teacher, enabling effective generalisation in real-world noisy scenarios.

Related Work

Adaptive Layer Attention: Transformer architectures in ASR often rely on final-layer encoder outputs, discarding intermediate representations that may carry phonetic, lexical, or semantic cues. Prior modalities, such as audio-visual ASR, have explored fusing features across layers to improve performance in noisy environments. For instance, MLCA-AVSR applies multi-layer cross-attention fusion to combine audio and visual embeddings at different encoder depths, yielding robustness to noisy speech (Wang et al. 2024; Hentschel et al. 2024).

More broadly, multi-layer feature fusion has been shown to be beneficial in vision and language modelling. While not directly targeting ASR, these studies motivate a dynamic combination of internal layer representations. Our ALA extends this principle within a single-modality ASR encoder. Although recently proposed Differential Transformer architectures introduce mechanisms to reduce attention to noise and mitigate hallucinations by amplifying relevant attention via subtraction across attention maps, these operate at the decoder head level rather than selectively fusing encoder layer representations (Ye et al. 2024). ALA complements such efforts by enriching encoder inputs before decoding.

Multi-Objective Knowledge Distillation: Knowledge distillation (KD) has been widely used to transfer robust behaviour from one model to another, often through hidden states matching. In speech recognition, KD has improved streaming or resource-constrained ASR by aligning CTC-based teacher state alignments or token posterior outputs to a student model (Hentschel et al. 2024).

In neural machine translation, attention alignment distillation techniques, such as "Align-to-Distill" (A2D), learn adaptive mappings between teacher and student attention heads, surpassing heuristic layer alignment and significantly boosting performance in low-resource scenarios (Jin et al. 2024; Inaguma and Kawahara 2021). Our MOKD framework adapts these insights to ASR: we align encoder and decoder final-layer representations and cross-attention between a clean-speech teacher and a noisy-speech student. Our losses include encoder cosine similarity, decoder cosine similarity, and decoder cross-attention map MSE, combined with standard token-level cross-entropy. This exhaustive supervision allows the student to replicate the teacher's robust attention behaviour and prevent hallucinations, going beyond traditional logit-only distillation. Prior works such as "Keep decoding parallel" show that intermediate-layer KD, particularly between an external language model and ASR student, can boost token accuracy in CTC-based or attention-based systems by transferring semantic LM knowledge at multiple levels (Jin et al. 2024; Hentschel et al. 2024). Similarly, our method leverages multi-layer internal alignment, but specifically targets hallucination reduction by mirroring the attention behaviour of a clean-trained teacher.

In Distil-Whisper (Gandhi, Von Platen, and Rush 2023), the authors used a simple word error rate (WER) heuristic to select only the highest quality pseudolabels for training. The study notes that Distil-Whisper is less prone to hallucination errors compared to the original Whisper model. Motivated by these findings, we use Distil-Whisper as one of the baseline models for this proposed approach. In combination, ALA and KD address hallucination in Whisper from two complementary angles: ALA enhances encoder representations by adaptively fusing layer blocks, improving contextual robustness under noise; KD aligns student decoder semantics and attention with that of a clean teacher. While previous literature has studied layer-wise fusion or attention alignment in isolation, our approach uniquely integrates both, yielding stronger WER reduction and hallucination mitigation in noisy ASR tasks.

Proposed Methodology

This section presents the key components of our methodology: (i) Adaptive Layer Attention (ALA), which constitutes Stage-1 and enhances encoder representations, and (ii) a Multi-Objective Knowledge Distillation (MOKD) framework, which forms Stage-2 and mitigates decoder hallucinations under noisy conditions.

Adaptive Layer Attention Details

Transformer-based ASR models like Whisper utilise deep encoder stacks, where each layer progressively abstracts fea-

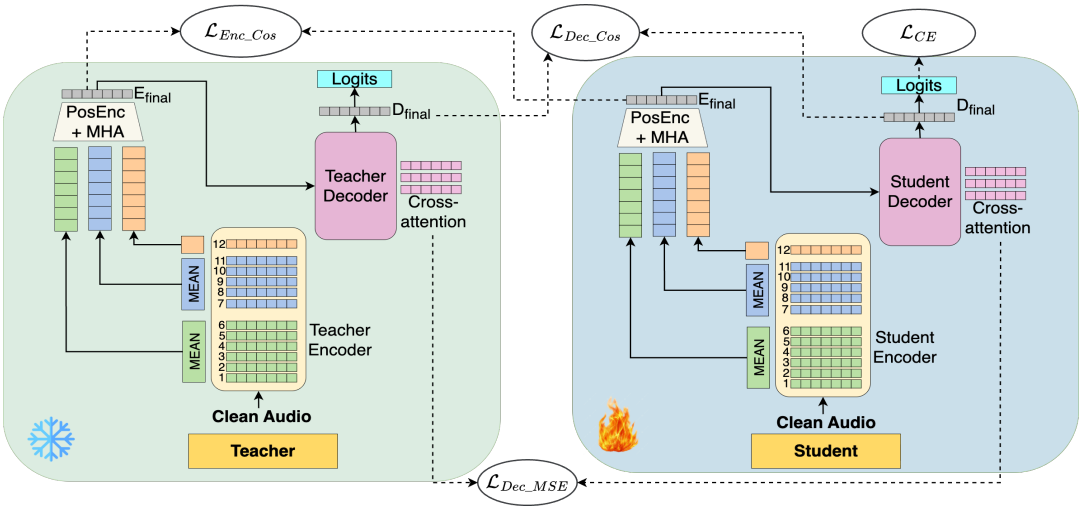


Figure 1: Block diagram of the proposed architecture combining Adaptive Layer Attention for encoder feature fusion and Multi-Objective Knowledge Distillation to reduce hallucination. E_{final} and D_{final} represents encoder and decoder final hidden states.

tures from raw audio to linguistic representations. However, under noisy conditions, certain encoder layers capture distorted or redundant signals, which degrade performance when passed directly to the decoder. To overcome this, we propose an ALA mechanism that adaptively fuses representations from structurally similar encoder layers, ensuring more robust and context-aware acoustic modelling.

Inter-Layer Similarity Analysis: We begin by analysing the cosine similarity between all encoder layer outputs to understand their functional roles under noise. The heatmap shown in Figure 2 reveals that layers tend to group naturally:

- Layers L1–L6 exhibit high mutual similarity, forming a block of low-level acoustic features.
- Layers L7–L11 form another block representing higher-level semantic abstractions.
- Layer L12 diverges significantly from all others, suggesting possible overfitting to noise and, notably, reflecting its specialised optimisation for decoder input.

This motivates a selective fusion strategy that avoids passing noisy or uninformative layers (e.g., L12) into the decoder, while still retaining useful abstractions. Notably, this structural pattern in encoder similarity was consistently observed across all languages evaluated in the dataset section.

Adaptive Fusion via Block-Wise Attention: Let the encoder produce hidden states from all L layers:

$$E = \{e_1, e_2, ..., e_L\}$$

We first compute pairwise cosine similarity between encoder layer outputs and group layers into K coherent blocks $\{B_1, B_2, ..., B_K\}$, where each block contains layers with high inter-layer similarity.

1. **Mean Block Representation:** To derive block-level representations, we explored several strategies, including weighted summation, multi-head attention, and mean pooling (Gwak and Jung 2025). Among these, mean

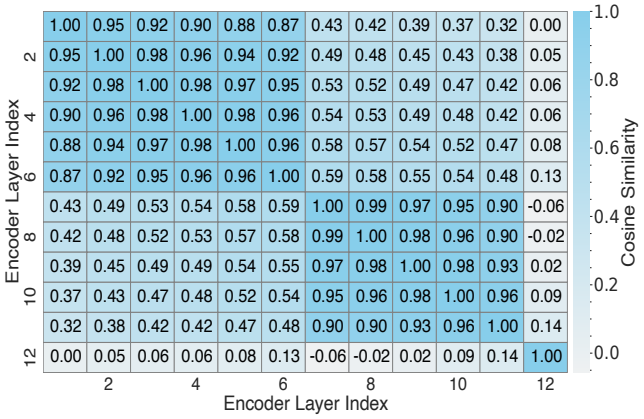


Figure 2: Heatmap for encoder inter-layer similarity on a subset of Hindi Kathbath test dataset.

pooling yielded the best performance. Consequently, for each block B_k, we compute a block-level representation r_k by applying the mean pooling to the hidden states of all layers in that block:

$$r_k = \frac{1}{|B_k|} \sum_{l \in B_k} e_l \quad (1)$$

This results in a set of block-wise embeddings $R = \{r_1, r_2, ..., r_K\}$.

2. **Positional Encoding:** To maintain temporal structure, we inject positional encoding (PosEnc) (Vaswani et al. 2017) into the mean block representations:

$$Z = PosEnc(R) \quad (2)$$

The resulting Z will then be passed to the MHA.

3. **Adaptive Multi-Head Attention and Final Projection:** For each token, we use the final encoder layer's hidden

state as the query and attend over the block representations Z using MHA.

$$h_t = \text{MHA}(q_t, Z, Z) \tag{3}$$

where q_t is the projected hidden state of the final encoder layer at position t. The attention output is then projected, added residually to the original query, and normalised. The resulting sequence $H = \{h_1, h_2, ..., h_T\}$ is passed to the decoder for prediction.

Multi-Objective Knowledge Distillation Details

Once the encoder is enhanced with ALA, we further strengthen decoder robustness using a student-teacher KD strategy. A clean-teacher model guides a noisy-student model, aligning their encoder and decoder representations. Using a noisy teacher resulted in inferior performance, confirming the effectiveness of clean supervision for robust ASR. The distillation framework uses multiple objectives to align the student's encoder and decoder with the clean teacher's representations.

Let (x^T, y^T) and (x^S, y^S) denote the clean teacher and noisy student input-output pairs, respectively. Here, e_t^T and e_t^S are the encoder hidden states at timsestep t, while d_t^T and d_t^S are the decoder hidden states at timestep t for the teacher and student models.

Loss Functions: The student model is optimised with a combination of four objectives:

1. **Encoder Cosine Similarity Loss:** We encourage alignment between teacher and student encoder representations using cosine similarity across the last layer:

$$\mathcal{L}_{Enc\_Cos} = \sum_{t=1}^{T} \left(1 - \cos\left(e_t^T, e_t^S\right)\right) \tag{4}$$

2. **Decoder Cosine Similarity Loss:** To match the contextual embedding space at the decoder level, we use cosine similarity on the last layer of the decoder:

$$\mathcal{L}_{Dec\_Cos} = \sum_{t=1}^{T} \left(1 - \cos\left(d_t^T, d_t^S\right)\right) \tag{5}$$

3. **Decoder Mean Squared Error (MSE) Loss:** We also apply MSE loss between the decoder cross-attention maps of the teacher and student:

$$\mathcal{L}_{Dec\_MSE} = \sum_{t=1}^{T} \left\| d_t^T - d_t^S \right\|_2^2 \tag{6}$$

4. **Cross-Entropy (CE) Loss:** The standard CE loss is used between the predicted token probabilities and the ground truth transcript:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \log P_S(y_t) \tag{7}$$

5. **Total Knowledge Distillation Loss:** The final loss combines all four components using tunable weighting coefficients:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Enc\_Cos}} + \lambda_2 \mathcal{L}_{\text{Dec\_Cos}} \\ + \lambda_3 \mathcal{L}_{\text{Dec\_MSE}} + \lambda_4 \mathcal{L}_{\text{CE}}\end{aligned} \tag{8}$$

We performed a grid search over the weighting coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ (ranging from 0.5 to 2.0) revealed that setting $\lambda_1 = 0.8$ and the others to 1.0 produced the best results. This multi-objective KD setup ensures that the student model not only learns to mimic the teacher's output but also captures deeper structural similarities in encoder and decoder spaces. This leads to reduced hallucination, especially in noisy speech scenarios, where decoder misalignment is more likely.

Experimental Setup

Implementation Details

All experiments were conducted on NVIDIA H100 GPUs. For ALA, we used 1 GPU, and MOKD was trained in a distributed fashion using 4 GPUs. Prior work (Atwany et al. 2025) highlights that smaller Whisper variants like whisper-tiny and Whisper-small exhibit higher word error rates and hallucination error rates, whereas larger models such as Whisper-medium and Whisper-large show notable improvements in performance. Based on this observation, we selected Whisper-small, referred to as W-S, as the base ASR model for all fine-tuning tasks in this study. For the proposed ALA approach, we fine-tuned the W-S model for 15 epochs using a learning rate of 5×10^{-5} for the base parameters and 9×10^{-5} for the parameters of the ALA module. The training included a warm-up phase of 5000 steps, followed by a cosine decay learning rate scheduler and 6 attention heads. For the MOKD approach, fine-tuning was performed for 72,000 steps using a learning rate of 1×10^{-5}, with 100 warm-up steps, and a linear learning rate decay schedule.

Dataset Details

We conducted experiments across Hindi, Arabic, French, and English to evaluate the robustness, generalizability, and effectiveness of our approach across diverse phonetic, syntactic, and noise-resilience characteristics. For Hindi ASR, we utilise the Kathbath train and test splits (Javed et al. 2023). Arabic and French models are trained and evaluated using the CommonVoice-15 dataset (Ardila et al. 2019), while for English ASR, we use LibriSpeech-100 for training and the test-clean split for evaluation (Panayotov et al. 2015). To replicate real-world noisy conditions and assess hallucination behaviour, we introduce noise from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) (Thiemann, Ito, and Vincent 2013). This dataset includes recordings from 18 varied acoustic settings, ranging from quiet indoor environments to noisy outdoor locations, captured using a 16-channel microphone array.

Audio files exceeding 30 seconds are segmented into chunks no longer than 30 seconds. For each language, we construct noisy training datasets with uniformly distributed signal-to-noise ratios (SNRs) ranging from -8 dB to +4 dB, and corresponding test sets with SNRs from -10 dB to +10 dB. Following (Zusag, Wagner, and Thallinger 2024), we also include noise-only samples (without speech) labelled with empty transcripts in 1% of the training data to further mitigate hallucinations during training.

Language	Model	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10	Clean	Average
Arabic	Baseline-1	259.59/0.0787	233.36/0.0803	235.17/0.0818	221.85/0.0820	220.07/0.0828	213.17/0.0836	230.53/0.0815
	Baseline-2	85.80/0.6421	69.68/0.7562	61.27/0.8226	57.20/0.8551	54.34/0.8739	52.26/0.8998	63.42/0.8083
	W-ALA	**77.51/0.6936**	**65.88/0.8008**	**58.20/0.8572**	**54.02/0.8863**	**52.06/0.8989**	**48.97/0.9118**	**59.44/0.8415**
French	Baseline-1	159.87/0.2912	142.05/0.3370	125.96/0.3715	116.65/0.3941	112.65/0.4049	110.44/0.4190	127.94/0.3696
	Baseline-2	64.12/0.7583	39.29/0.8521	28.82/0.8994	24.42/0.9208	22.06/0.9312	19.76/0.9405	33.08/0.8837
	W-ALA	**57.26/0.7804**	**37.15/0.8645**	**27.23/0.9076**	**23.13/0.9263**	**20.81/0.9354**	**18.91/0.9438**	**30.75/0.8930**
Hindi	Baseline-1	158.00/0.0897	149.53/0.0919	145.90/0.0915	132.82/0.0914	127.86/0.0911	132.38/0.0911	141.08/0.0911
	Baseline-2	42.77/0.8027	26.65/0.8946	18.05/0.9365	14.95/0.9512	13.44/0.9580	12.77/0.9638	21.44/0.9178
	W-ALA	**40.74/0.8257**	**24.15/0.9107**	**16.07/0.9448**	**13.33/0.9588**	**12.13/0.9636**	**11.41/0.9669**	**19.64/0.9284**
English	Baseline-1	40.27/0.8252	16.20/0.9237	7.11/0.9648	5.55/0.9789	4.91/0.9827	3.82/0.9853	12.97/0.9434
	Baseline-2	39.64/0.8763	16.02/0.9434	7.21/0.9711	4.56/0.9810	3.91/0.9837	3.44/0.9851	12.46/0.9567
	W-ALA	**29.68/0.8772**	**11.93/0.9450**	**5.85/0.9727**	**3.98/0.9823**	**3.45/0.9854**	**3.19/0.9866**	**9.68/0.9581**

Table 1: Comparison of Baseline-1, Baseline-2 and W-ALA models across various noise levels and clean audio for four languages. Each cell shows WER (↓) / SeMaScore (↑).

Comparison Methods

For the first stage we apply Adaptive Layer Attention (ALA) to the Whisper encoder to improve its robustness under noisy conditions. The second stage builds on this by introducing a multi-objective Knowledge Distillation (MOKD) strategy to further reduce hallucinations and enhance transcription quality. For comparison, we evaluate three baselines: (1) the original pre-trained Whisper-small model from OpenAI, which we call Baseline-1, (2) Whisper-small fine-tuned on our noisy training set, Baseline-2, and (3) Distil-Whisper (Gandhi, Von Platen, and Rush 2023), where the fine-tuned model from Baseline-2 serves as the teacher and a distilled student model retains all encoder layers but only two decoder layers (initialized from the teacher's first and last decoder layers), with the encoder kept frozen. We call this Baseline-3. To ensure a fair comparison, all baselines and our proposed methods are trained and evaluated on the same noisy dataset (SNR -8 to +4 dB) described in the Dataset Details section.

Evaluation Metrics

ASR performance is conventionally evaluated using Word Error Rate (WER in %). However, as highlighted in (Sasindran, Yelchuri, and Prabhakar 2024), WER has notable limitations as it fails to capture semantic similarity and does not account for the varying importance of different word errors. To address this, recent works (Kim et al. 2021a,b; Whetten and Kennington 2023; Sasindran et al. 2023) have employed pretrained BERT-based models (Devlin et al. 2019) to estimate the semantic distance between the reference and hypothesis. Yet, sentence-level semantic metrics can disproportionately weight certain words, reducing their reliability. In contrast, SeMaScore (Sasindran, Yelchuri, and Prabhakar 2024) remains effective even when WER is high, particularly under challenging conditions such as noisy speech. Accordingly, this work also evaluates hallucination in ASR outputs using SeMaScore, which ranges between 0 and 1.

Model	Latency (ms)	RTF	Peak VRAM
Baseline-2	140 ± 10	0.021	1.5
W-ALA	152 ± 11	0.023	2.6

Table 2: Latency, RTF, and peak memory usage (in GB) comparison between Baseline-2 and W-ALA models.

Results and Discussions

In this work, we conducted different evaluations for stage-1 and stage-2 approaches.

Stage-1 Evaluation

Table 1 presents the first stage of evaluation, where we examine the effectiveness of the ALA mechanism across four languages: Arabic, French, Hindi, and English, under a range of noise conditions, from –10 dB SNR to clean speech. Compared to both Baseline-1 and Baseline-2, the Whisper finetuned with ALA, referred to as the W-ALA model, consistently achieves lower WER and higher SeMaScore, demonstrating enhanced robustness and semantic preservation in noisy environments. Notably, proposed W-ALA offers significant improvements at low SNRs (e.g., -10 dB, 5dB and 0 dB), where conventional baselines degrade sharply. On average, W-ALA reduces WER by a substantial margin while simultaneously increasing cosine similarity, indicating more reliable and confident predictions. These results validate the ALA module's ability to dynamically prioritise acoustically informative encoder layers, improving recognition performance without compromising semantic correctness.

We benchmarked the inference efficiency of the W-ALA model against Baseline-2 on the LibriSpeech English test set using a batch size of 1. As shown in Table 2 , W-ALA adds only minimal overhead: latency increases by 8%, real-time factor (RTF) by 9%. Peak VRAM usage rises by approximately 1 GB, remaining well within the capacity of modern accelerators. Despite introducing just 0.98% more parameters, W-ALA delivers significant gains in robustness and accuracy (see Table 1) over Baseline-2. These results

Language	Model	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10	Clean	Average
Arabic	Baseline-2	39.64/0.8763	16.02/0.9434	7.21/0.9711	4.56/0.9810	3.91/0.9837	3.44/0.9851	12.46/0.9567
	Baseline-3	83.53/0.6844	75.95/0.6038	69.77/0.6734	65.81/0.7176	63.03/0.7412	60.33/0.7628	69.74/0.6639
	W-MOKD	**76.32/0.7126**	**62.85/0.8021**	**55.50/0.8719**	**51.53/0.8973**	**49.97/0.9063**	**48.21/0.9202**	**57.40/0.8557**
French	Baseline-2	64.12/0.7583	39.29/0.8521	28.82/0.8994	24.42/0.9208	22.06/0.9312	19.76/0.9405	33.08/0.8837
	Baseline-3	63.60/0.6988	47.12/0.7877	38.98/0.8301	34.02/0.8515	31.64/0.8616	28.23/0.8796	40.60/0.8181
	W-MOKD	**51.41/0.7690**	**34.42 / 0.8767**	**25.44/0.9152**	**21.40/0.9318**	**19.61/0.9393**	**17.79/0.9468**	**28.18/0.8965**
Hindi	Baseline-2	42.77/0.8027	26.65/0.8946	18.05/0.9365	14.95/0.9512	13.44/0.9580	12.77/0.9638	21.44/0.9178
	Baseline-3	56.16/0.6950	54.72/0.7412	48.20/0.7755	46.53/0.7757	46.28/0.7749	47.43/0.7602	51.39/0.7544
	W-MOKD	**38.13/0.8455**	**22.67/0.9257**	**14.83/0.9580**	**12.97/0.9697**	**11.86/0.9644**	**11.23/0.9675**	**18.61/0.9432**
English	Baseline-2	39.64/0.8763	16.02/0.9434	7.21/0.9711	4.56/0.9810	3.91/0.9837	3.44/0.9851	12.46/0.9567
	Baseline-3	69.75/0.7634	55.79/0.8327	44.69/0.8660	45.36/0.8892	42.47/0.8767	44.11/0.9029	50.36/0.8592
	W-MOKD	**26.43/0.9046**	**9.41/0.9726**	**5.72/0.9842**	**3.11/0.9853**	**3.49/0.9832**	**3.18/0.9836**	**8.56/0.9690**

Table 3: Comparison of Baseline-2, Baseline-3 and W-MOKD models across various noise levels and clean audio for four languages. Each cell shows WER (↓) / SeMaScore (↑).

highlight that W-ALA achieves improved performance with negligible impact on runtime, memory usage and additional parameters, making it a practical and scalable solution for real-world ASR applications.

Stage-2 Evaluation

Table 3 presents a comprehensive evaluation of the proposed W-MOKD model across four languages: Arabic, French, Hindi, and English, under a range of noisy (SNR-10 to 10 dB) and clean conditions. The results demonstrate that W-MOKD consistently outperforms Baseline-2 and the strong Distil-Whisper baseline (Baseline-3) in both WER and SeMaScore across all languages and SNR levels. Notably, the gains are more significant under severe noise (e.g., at -10 dB). Even under clean conditions, W-MOKD yields improvements in semantic accuracy. Moreover, W-MOKD outperforms the W-ALA model. These results validate the effectiveness of our two-stage architecture in enhancing robustness and reducing hallucinations across diverse linguistic and acoustic settings, particularly under challenging noise conditions.

The average WER and SeMaScore across noise levels show clear gains, highlighting that the proposed MOKD framework effectively suppresses hallucinations and enhances semantic accuracy, particularly in challenging acoustic conditions. Overall, these results validate the effectiveness of combining adaptive encoder attention with attention- and representation-level distillation, resulting in more noise-robust and semantically accurate ASR across multiple languages.

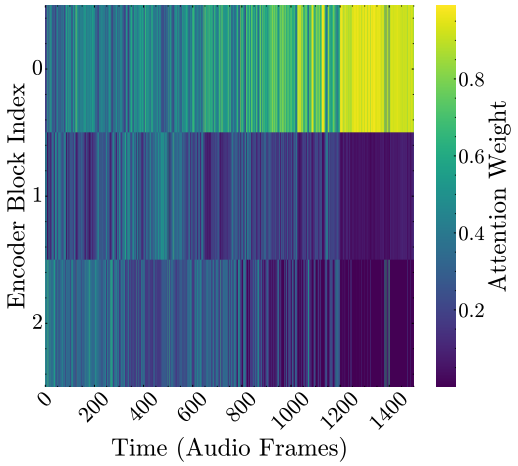
Analysing Encoder Block Robustness

Figure 3 presents an analysis of encoder block attention behaviour under noisy conditions, highlighting the robustness of different blocks in the ALA mechanism. The heatmap (Figure 3a) shows frame-wise attention distribution for a noisy audio sample, where Block 0 consistently receives the highest attention, especially during segments likely impacted by noise. This indicates that Block 0 captures more stable and noise-resilient features during inference. In contrast, Blocks 1 and 2 are less frequently attended, suggesting the model relies less on deeper or more abstract representations in noisy scenarios. The accompanying bar chart (Figure 3b) shows average attention weights across the entire dataset, reinforcing this pattern: Block 0 dominates with an average weight of 0.586, while Blocks 1 and 2 receive significantly lower and nearly equal attention. Together, these results demonstrate a consistent model preference for Block 0 in noisy settings. This validates the effectiveness of Adaptive Layer Attention (ALA) in dynamically emphasising robust encoder features for improved ASR performance.

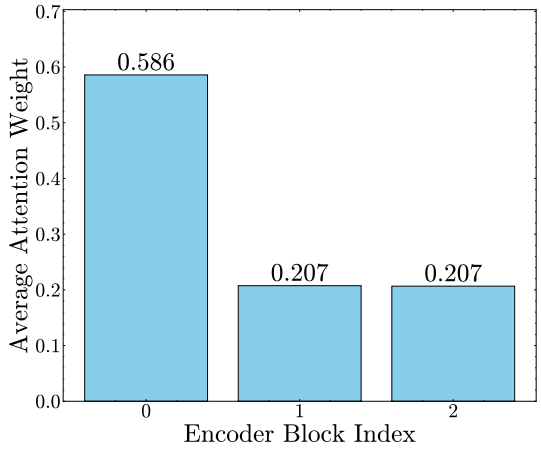
Stage-1 Ablation Study

In Table 4, the Stage-1 ablation study investigates various strategies for fusing encoder representations in Whisper using ALA, conducted on the Hindi dataset at –10 dB SNR and clean. The first method, a weighted sum across all encoder layers, performs the worst (WER: 75.85/29.56, SeMaScore: 0.4822/0.7521) on -10 dB and clean, likely due to uniformly treating layers without capturing their functional diversity. Applying MHA over all encoder layers while keeping them frozen yields a slightly better WER (52.73/15.62), indicating frozen layers cannot adapt to noise. When layers are trainable during MHA, performance improves significantly (WER: 45.12/14.87, SeMaScore: 0.6902/0.9290), demonstrating that encoder layers begin learning noise-robust features.

We further group encoder layers into three blocks based on inter-layer similarity and experiment with MHA on first (1,7,12), last (6,11,12), and middle (3,9,12) layers of each block. These configurations yield improved WERs and better SeMaScores. The best results (WER: 40.74/11.41, SeMaScore: 0.8257/0.9669) come from applying MHA over mean representations from each block, highlighting the benefit of block-level context aggregation for robust and hallucination-resistant ASR.



(a) Block attention across time for a sample



(b) Dataset-wide average attention weights

Figure 3: Layer-wise attention behaviour for noisy speech input.

Fusion Method	SNR -10	Clean
Weighted Sum	75.85/0.4822	29.56/0.7521
MHA all frozen	52.73/0.5454	15.62/0.8928
MHA all	45.12/0.6902	14.87/0.9290
MHA 1,7,12	42.23/0.7032	14.20/0.9352
MHA 6,11,12	41.91/0.7105	12.16/0.9480
MHA 3,9,12	41.53/0.7135	12.14/0.9492
MHA Mean	**40.74/0.8257**	**11.41/0.9669**

Table 4: Comparison of encoder-fusion strategies under noisy (-10 dB) and clean conditions on the Hindi test set.

KD Loss Function	SNR -10	Clean
CE (no KD)	42.77/0.8027	12.77/0.9638
+ KL_{logits}	46.48/0.7331	14.61/0.9413
+ CosAll	45.61/0.8071	14.61/0.9592
+ CosFin	**42.97/0.8246**	**12.37/0.9622**
+ OTFin	46.53/0.8019	13.56/0.9614
+ IRLEFin	61.14/0.6102	39.36/0.7333
+ KLFin	43.88/0.8193	13.15/0.9625
+ CosFin+MSE_{decCA}	42.46/0.7207	12.52/0.9486
+ CosFin+MSE_{decCA}+MSE_{decSA}	42.82/0.7192	12.97/0.9373
+ CosFin+MSE_{decCA}+MSE_{enc}	42.73/0.7137	12.94/0.9345
+ CosFin+MSE_{decCA}+ALA	**38.13/0.8455**	**11.23/0.9675**

Table 5: Comparison of KD loss functions under noisy (-10 dB) and clean conditions.

Stage-2 Ablation Study

Table 5 presents an ablation study assessing the impact of different KD loss functions under both -10 dB and clean conditions. Here, CosAll applies cosine similarity loss across all encoder and decoder hidden states, while CosFin restricts it to only the final layers. Similarly, Optimal Transport OTFin, IRLEFin, and KLFin apply OT (Mao, Chen, and Li 2025), IRLE (Higuchi et al. 2021), and KL divergence losses to final encoder and decoder representations. MSE_{decCA} denotes mean squared error loss between decoder cross-attention maps, and MSE_{decSA} is between decoder self-attention maps, and MSE_{enc} represents loss between encoder attention maps.

Among the methods, CosFin offers a strong trade-off, showing that deep-layer alignment is more semantically meaningful, performing better alignment over CosAll. KLFin provide marginal gains, while distance-based losses like IRLE and OT significantly degrade performance. Adding only MSE_{decCA} offers the best improvement across all attention losses. The best results emerge when combining CosFin and MSE_{decCA} with ALA.

Conclusions

In this work, we proposed a two-stage framework to improve the robustness and semantic fidelity of Whisper model under noisy conditions. In Stage-1, we introduced an adaptive layer attention mechanism that dynamically fuses encoder representations based on inter-layer similarity. This approach enables the model to emphasise the most informative and noise-resilient encoder blocks, effectively suppressing irrelevant activations that often lead to hallucinations. In Stage-2, we enhanced this with a multi-objective knowledge distillation framework that aligns the student model's encoder representations, decoder semantics, and cross-attention maps with those of a clean teacher. Our experiments across multiple languages and varying noise levels demonstrate substantial improvements in both word error rate and SeMaScore, particularly under challenging low-SNR conditions. Overall, our framework offers a principled approach to mitigate hallucinations and improve Whisper reliability in real-world noisy environments. Future work can explore cross-lingual generalisation and apply our proposed approach to other transformer-based speech models.

References

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Atwany, H.; Waheed, A.; Singh, R.; Choudhury, M.; and Raj, B. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. *arXiv preprint arXiv:2502.12414*.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Barakat, O. T. . C. D., Huda. 2024. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 11(1).

Barański, M.; Jasiński, J.; Bartolewska, J.; Kacprzak, S.; Witkowski, M.; and Kowalczyk, K. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Fang, Z.; Zhang, R.; He, Z.; Wu, H.; and Cao, Y. 2022. Non-autoregressive Chinese ASR error correction with phonological training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5907–5917.

Frieske, R.; and Shi, B. E. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*.

Gandhi, S.; Von Platen, P.; and Rush, A. M. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Gwak, J.; and Jung, Y. 2025. Layer-Aware Embedding Fusion for LLMs in Text Classifications. *arXiv preprint arXiv:2504.05764*.

Hentschel, M.; Nishikawa, Y.; Komatsu, T.; and Fujita, Y. 2024. Keep decoding parallel with effective knowledge distillation from language models to end-to-end speech recognisers. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10876–10880. IEEE.

Higuchi, Y.; Tawara, N.; Ogawa, A.; Iwata, T.; Kobayashi, T.; and Ogawa, T. 2021. Noise-robust attention learning for end-to-end speech recognition. In *2020 28th European Signal Processing Conference (EUSIPCO)*, 311–315. IEEE.

Inaguma, H.; and Kawahara, T. 2021. Alignment knowledge distillation for online streaming attention-based speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1371–1385.

Javed, T.; Bhogale, K.; Raman, A.; Kumar, P.; Kunchukuttan, A.; and Khapra, M. M. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12942–12950.

Jin, H.; Son, S.; Park, J.; Kim, Y.; Noh, H.; and Lee, Y. 2024. Align-to-Distill: Trainable Attention Alignment for Knowledge Distillation in Neural Machine Translation. *arXiv preprint arXiv:2403.01479*.

Kim, S.; Arora, A.; Le, D.; Yeh, C.-F.; Fuegen, C.; Kalinli, O.; and Seltzer, M. L. 2021a. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv preprint arXiv:2104.02138*.

Kim, S.; Le, D.; Zheng, W.; Singh, T.; Arora, A.; Zhai, X.; Fuegen, C.; Kalinli, O.; and Seltzer, M. L. 2021b. Evaluating user perception of speech recognition system quality with semantic distance metric. *arXiv preprint arXiv:2110.05376*.

Mao, S.; Chen, R.; and Li, H. 2025. Weighted joint distribution optimal transport based domain adaptation for cross-scenario face anti-spoofing. *International Journal of Computer Vision*, 133(2): 590–610.

Nguyen, H.; He, Z.; Gandre, S. A.; Pasupulety, U.; Shivakumar, S. K.; and Lerman, K. 2025. Smoothing Out Hallucinations: Mitigating LLM Hallucination with Smoothed Knowledge Distillation. *arXiv preprint arXiv:2502.11306*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Sasindran, Z.; Yelchuri, H.; and Prabhakar, T. 2024. SeMaScore: a new evaluation metric for automatic speech recognition tasks. *arXiv preprint arXiv:2401.07506*.

Sasindran, Z.; Yelchuri, H.; Prabhakar, T.; and Rao, S. 2023. H eval: A new hybrid evaluation metric for automatic speech recognition tasks. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–7. IEEE.

Thiemann, J.; Ito, N.; and Vincent, E. 2013. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19, 035081. Acoustical Society of America.

Tseng, L.-H.; Chen, Z.-C.; Chang, W.-S.; Lee, C.-K.; Huang, T.-R.; and Lee, H.-y. 2024. Leave no knowledge behind during knowledge distillation: Towards practical and effective knowledge distillation for code-switching asr using realistic data. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 118–125. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Guo, P.; Zhou, P.; and Xie, L. 2024. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8150–8154. IEEE.

Whetten, R.; and Kennington, C. 2023. Evaluating and improving automatic speech recognition using severity. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 79–91.

Yang, R.; Wu, T.; Wang, J.; Hu, P.; Wu, Y.-C.; Wong, N.; and Yang, Y. 2024. Llm-neo: Parameter efficient knowledge distillation for large language models. *arXiv preprint arXiv:2411.06839*.

Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; and Wei, F. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Yoon, J. W. 2025. Heuristic-free Knowledge Distillation for Streaming ASR via Multi-modal Training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25733–25741.

Zusag, M.; Wagner, L.; and Thallinger, B. 2024. Crisper-whisper: Accurate timestamps on verbatim speech transcriptions. In *Proc. Interspeech 2024*, 1265–1269.