MODALITY-ADAPTIVE DEPTH UP-SCALING FOR TRIMODAL LLM ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting pre-trained text large language models (LLMs) to new modalities via continual pre-training often degrades their text capabilities. Depth up-scaling, which inserts new layers into a pre-trained LLM, has gained attention for adding capacity while preserving the original capabilities, and has been shown to be effective for adapting text LLMs to speech. We present *modality-adaptive depth up-scaling*, which extends this approach to a trimodal (text, speech, and image) setting, in which a single text LLM is adapted to both speech and image through shared expansion layers. Since speech and image tokens form a 1D temporal sequence and a 2D spatial grid, respectively, we introduce modality-adaptive (MA) operations, namely 1D convolution for speech tokens and 2D convolution for image tokens, and build each expansion layer as an MA-Conformer layer that switches between them by token type. We also introduce Expansion LayerDrop, which lets a single trained model trade performance for speed by activating fewer expansion layers at inference. On speech-to-text, image-to-text, and jointly trained trimodal adaptation with 1.7B and 3B base LLMs, MA-Conformer expansion layers outperform standard transformer expansion layers and cause less text degradation than full fine-tuning, and a single trimodal model approaches the performance of separately trained unimodal models.

1 INTRODUCTION

Text LLMs have been extended to speech (Chu et al., 2024; Tian et al., 2025) and vision (Liu et al., 2024a) via continual pre-training (CPT), but CPT on new-modality data degrades their text capabilities (Hsiao et al., 2025; Zhang et al., 2024b). Replaying the original pre-training data is not possible for most open-weight LLMs, which do not release it (Grattafiori et al., 2024; Yang et al., 2025a), and parameter-efficient fine-tuning (Hu et al., 2022) may lack the capacity (Biderman et al., 2024) for an entirely new modality. Depth up-scaling inserts new layers into a pre-trained LLM (Kim et al., 2024; Wu et al., 2024), and keeping the original layers frozen avoids both limitations while allowing the pre-trained model to be recovered exactly by removing the new layers. Yano et al. (2026) applied this approach to speech, reaching automatic speech recognition (ASR) performance comparable to full fine-tuning with far less text degradation. However, that study covered only speech, leaving open whether a single model can be adapted to several modalities.

In this work, we present *modality-adaptive depth up-scaling*, which extends the depth up-scaling approach of Yano et al. (2026) from speech to a trimodal (text, speech, and image) setting, in which a single text LLM is adapted to both speech and image through the same expansion layers (Figure 1). Since speech and image tokens form a 1D temporal sequence and a 2D spatial grid, respectively, we introduce modality-adaptive (MA) operations, namely 1D convolution for speech tokens and 2D convolution for image tokens, and build each expansion layer as an MA-Conformer layer, which follows the Conformer (Gulati et al., 2020) and switches between these operations by token type.

We evaluate this framework on speech-to-text, the setting of Yano et al. (2026), and on two settings beyond it, namely image-to-text and jointly trained trimodal adaptation, using SmolLM2-1.7B (Ben Allal et al., 2025) and SmolLM3-3B-Base (Bakouch et al., 2025) as base LLMs. In all settings, our central proposal, matching the local operation of the expansion layers to the structure of each modality, improves adaptation over standard transformer expansion layers, and modality-adaptive depth up-

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

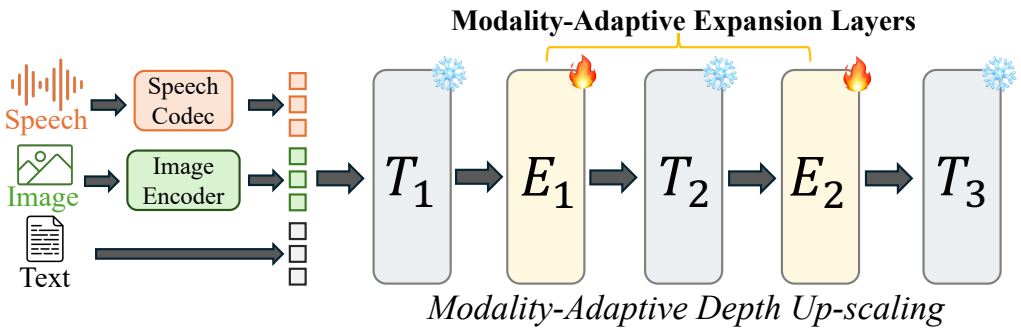

Figure 1: Trainable expansion layers E are interleaved with frozen original layers T of a pre-trained text LLM. The expansion layers apply modality-adaptive modules to speech and image tokens. For text-only deployment, removing the expansion layers (and any added vocabulary entries) exactly recovers the original text model.

scaling reaches task performance comparable to full fine-tuning with smaller text degradation. A single trimodal model approaches separately trained unimodal models.

Our contributions are as follows. **(1)** We extend depth up-scaling, previously applied to text and to speech (Yano et al., 2026), to image and to jointly trained trimodal adaptation. **(2)** We introduce MA operations and the MA-Conformer, which switches between 1D convolution for speech tokens and 2D convolution for image tokens by input token type. **(3)** We introduce Expansion LayerDrop, which applies structured dropout (Huang et al., 2016; Fan et al., 2020) only to the expansion layers over a frozen base model, unlike prior layer dropping (Fan et al., 2020; Elhoushi et al., 2024), so that a single model supports multiple performance-speed operating points.

2 Method

Our goal is to adapt a pre-trained text LLM to accept multimodal input while keeping its original text capabilities exactly recoverable by design. We introduce a depth up-scaling framework (Section 2.1) with modality-adaptive Conformer (MA-Conformer) expansion layers (Section 2.2), instantiated for speech, image, and trimodal settings (Section 2.3), and Expansion LayerDrop (Section 2.4).

2.1 Depth Up-Scaling for Multimodal Adaptation

We consider the task of adapting a pre-trained text LLM to accept multimodal input (speech, image, or both) and produce text output via the same autoregressive generation process as the base model. The base model has L transformer layers $T_1, \ldots, T_L$. Given an input embedding $\mathbf{H}^{(0)} \in \mathbb{R}^{N \times d}$, where N is the sequence length and d is the hidden dimension, the standard forward pass computes $\mathbf{H}^{(l)} = T_l(\mathbf{H}^{(l-1)})$ for $l = 1, \ldots, L$.

To adapt the LLM to a new modality, we insert K additional *expansion layers* $E_1, \ldots, E_K$ at a fixed set of positions $\mathcal{S} = \{s_1, \ldots, s_K\}$ with $s_1 < \cdots < s_K$. The modified forward pass becomes:

$$\mathbf{H}^{(l)} = \begin{cases} E_j\left(\mathbf{H}^{(l-1)}\right) & \text{if } l \in \mathcal{S}, \\ T_{l'}\left(\mathbf{H}^{(l-1)}\right) & \text{otherwise,} \end{cases} \tag{1}$$

where l ranges over $1, \ldots, L+K$, j is the unique index such that $l = s_j$, and $l' = l - |\{s \in \mathcal{S} : s \leq l\}|$ indexes the corresponding original layer. The parameters of $\{T_l\}_{l=1}^{L}$ are frozen, and only $\{E_j\}_{j=1}^{K}$ and the modality interface (Section 2.3) are updated. The placement $\mathcal{S}$ affects both task performance and text preservation. It can distribute the expansion layers evenly (Interleaved), concentrate them in the lower or upper portions (Bottom, Top), or place them at both ends (Sandwich). We use Interleaved, which inserts one expansion layer after every L/K original layers (K divides L here) so that $s_j = j(L/K + 1)$, in all main experiments, as it gives the best task performance in prior work on speech (Yano et al., 2026) and in our image placement ablation (Appendix I).

Because the parameters of $\{T_l\}_{l=1}^{L}$ are never modified, removing all expansion layers $\{E_j\}_{j=1}^{K}$ at inference, that is, setting $K = 0$ in Equation (1) so that $\mathcal{S} = \emptyset$ and $l' = l$, recovers the standard forward pass $\mathbf{H}^{(l)} = T_l(\mathbf{H}^{(l-1)})$ exactly, guaranteeing zero text degradation for text-only deployment. In a practical multimodal deployment, however, text requests may also be served with the expansion layers active, for example when inputs mix modalities or when the expansion layers are not switched off per request. We therefore evaluate all text benchmarks with expansion layers active (the stricter criterion), and refer to text preservation in this setting as *active-mode retention* to distinguish it from the exact recoverability guarantee.

2.2 Modality-Adaptive Conformer

The choice of expansion layer architecture matters because multi-head self-attention (MHSA) alone, while effective at capturing global context, treats all positions uniformly and lacks the local inductive biases that modality-specific signals require. Convolution addresses this directly, capturing local temporal structure in speech (Amodei et al., 2016; Gulati et al., 2020; Kim et al., 2023) and local spatial structure in vision (Szegedy et al., 2015; He et al., 2016; Wu et al., 2021; Liu et al., 2022). Yet depth up-scaling for text-domain expansion has used standard transformer layers as the added layers (Kim et al., 2024; Wu et al., 2024; Du et al., 2024), identical to the original frozen layers. Since the expansion layers bear the full burden of adapting frozen LLM representations to new modalities, we expect architecture choice to matter more here than in text-domain expansion. We therefore adopt Conformer-style layers (Gulati et al., 2020), which have proven effective in speech (Gulati et al., 2020; Kim et al., 2023) by complementing MHSA with depthwise convolution[1].

We then introduce the MA-Conformer with two convolutional gating MLP (cgMLP) variants, whose full definitions are given in Appendix D. $\mathrm{cgMLP}_{\mathrm{1D}}$ applies 1D depthwise convolution suited for temporal speech sequences, and $\mathrm{cgMLP}_{\mathrm{2D}}$ reshapes the image token sequence into a 2D patch grid and applies 2D depthwise convolution to capture local spatial structure. The variant is selected at runtime by modality $m \in \{\texttt{speech}, \texttt{image}\}$:

$$\mathrm{cgMLP}^{(m)}(\cdot) = \begin{cases} \mathrm{cgMLP}_{\mathrm{1D}}(\cdot) & \text{if } m = \texttt{speech}, \\ \mathrm{cgMLP}_{\mathrm{2D}}(\cdot) & \text{if } m = \texttt{image}. \end{cases} \tag{2}$$

For an expansion layer E_j at position $l = s_j$ in Equation (1), we write its input $\mathbf{H}^{(l-1)}$ as $\mathbf{H} \in \mathbb{R}^{N \times d}$, omitting the layer index l. $\mathbf{H}$ consists of text tokens and the tokens of one or more modalities, and we write $\mathcal{I}_m \subseteq \{1, \ldots, N\}$ for the positions of the tokens of modality $m \in \{\texttt{text}, \texttt{speech}, \texttt{image}\}$ and $\mathbf{H}[\mathcal{I}_m]$ for the corresponding rows. The MA-Conformer layer computes $E_j(\mathbf{H})$ as

$$\mathbf{H}_1 = \mathbf{H} + \mathrm{MHSA}(\mathrm{RMSNorm}(\mathbf{H})), \tag{3}$$

$$\mathbf{H}_2[\mathcal{I}_m] = \begin{cases} \mathbf{H}_1[\mathcal{I}_m] + \mathrm{cgMLP}^{(m)}(\mathrm{RMSNorm}(\mathbf{H}_1[\mathcal{I}_m])) & \text{if } m \in \{\texttt{speech}, \texttt{image}\}, \\ \mathbf{H}_1[\mathcal{I}_m] & \text{if } m = \texttt{text}, \end{cases} \tag{4}$$

$$E_j(\mathbf{H}) = \mathbf{H}_2 + \mathrm{FFN}(\mathrm{RMSNorm}(\mathbf{H}_2)), \tag{5}$$

where Equation (4) is applied to each modality m present in the input. Equation (3) uses the same causal attention mask as the base LLM, so each token attends to all preceding tokens of any modality. All tokens are then processed by the same feed-forward network (FFN) in Equation (5).

2.3 Modality-Specific Instantiation

Our framework is agnostic to how modality tokens are produced, provided that image tokens keep their 2D patch-grid layout. In the *tokenizer-based* approach, a codec or neural tokenizer (van den Oord et al., 2017; Esser et al., 2021; Défossez et al., 2023) converts the raw signal into discrete tokens appended to the LLM vocabulary, as in OpusLM (Tian et al., 2025), and exact recovery (Section 2.1) then also removes these added entries. In the *encoder-based* approach, a pre-trained modality encoder such as CLIP (Radford et al., 2021) produces continuous embeddings projected into the LLM's hidden space via a learned projection, as in LLaVA (Liu et al., 2023) and its successors (Liu et al., 2024a; Li et al., 2025). Speech and image embeddings are concatenated into a single sequence, modality labels are assigned accordingly, and the expansion layers process them together.

[1]The original Conformer uses a macaron-style structure with two half-step FFNs before and after the main modules. We use a single FFN after the cgMLP for simplicity.

2.4 EXPANSION LAYERDROP

Removing all expansion layers recovers the text LLM (Section 2.1), but modality inputs pass through all K expansion layers, increasing per-token cost relative to the original model. Inference speed is a practical constraint when deploying multimodal models on resource-limited devices, in both speech recognition (Radford et al., 2023; Kim et al., 2022) and vision-language understanding (Liang et al., 2025b; Ye et al., 2025). We therefore introduce Expansion LayerDrop, which activates only a subset $\mathcal{A} \subseteq \mathcal{S}$ of $k = |\mathcal{A}|$ expansion layers at inference. An inactive expansion layer is skipped, so the first case of Equation (1) becomes $\mathbf{H}^{(l)} = E_j(\mathbf{H}^{(l-1)})$ if $l \in \mathcal{A}$ and $\mathbf{H}^{(l)} = \mathbf{H}^{(l-1)}$ if $l \in \mathcal{S} \setminus \mathcal{A}$. Setting $\mathcal{A} = \mathcal{S}$ gives the full model, and $\mathcal{A} = \emptyset$ recovers the original text LLM for text input. To make every $k \geq 1$ usable from a single model, we drop expansion layers during training (Huang et al., 2016; Fan et al., 2020) while the base LLM remains frozen, and study three schedules for sampling $\mathcal{A}$.

Uniform Drop. Following LayerDrop (Fan et al., 2020), each expansion layer is independently dropped with the same probability p during training. Unlike LayerDrop, which applies dropping to all transformer layers, we restrict it to the expansion layers only, leaving the frozen base LLM intact.

Linear Drop. Motivated by stochastic depth (Huang et al., 2016), the drop probability increases linearly from $p_{\min}$ for E_1 to $p_{\max}$ for E_K. This schedule places less pressure on early expansion layers, which must transform the input into a representation the frozen base layers can process. At inference, Uniform Drop samples k layers at random, and Linear Drop keeps the first k.

Ordered Drop. Uniform and Linear Drop spread training over many different subsets of expansion layers, so any particular subset used at inference with small k is rarely seen during training. We therefore introduce Ordered Drop, which activates only the first k expansion layers, $\mathcal{A} = \{s_1, \ldots, s_k\}$, in both training and inference. At each training step, k is sampled with probability proportional to $(k+1)$ for $k \in \{1, \ldots, K\}$, weighting larger k more frequently to preserve full-model performance while covering smaller configurations. Ordered Drop has been applied in the width direction for federated learning efficiency (Horváth et al., 2021) and in the depth direction by LayerSkip (Elhoushi et al., 2024), which combines layer dropping with an early-exit loss across all layers. Our method differs in dropping only the expansion layers over a frozen base LLM.

3 EXPERIMENTS

We evaluate modality-adaptive depth up-scaling in three settings. Speech-to-text (Section 3.1) follows the setting of Yano et al. (2026), where we confirm that MA-Conformer expansion layers work at least as well as the transformer expansion layers used there. Image-to-text (Section 3.2) applies the same framework to a new modality, with 2D instead of 1D convolution. Jointly trained trimodal adaptation (Section 3.3), the main extension of this work, adapts a single model to both speech and image. We then evaluate Expansion LayerDrop (Section 3.4).

Common setup. All settings use SmolLM2-1.7B (Ben Allal et al., 2025) (24 layers) and SmolLM3-3B-Base (Bakouch et al., 2025) (36 layers) as base LLMs, with K expansion layers equal to 25% of the original depth ($K = 6$ and $K = 9$) in INTERLEAVED placement (Section 2.1). We compare against full fine-tuning, which updates all original LLM parameters, and low-rank adaptation (LoRA) (Hu et al., 2022), whose rank r is set to match the trainable parameter count of depth up-scaling with standard transformer expansion layers[2]. Because the convolutional branch of the MA-Conformer adds parameters, at 1.7B we also train a parameter-matched MA-Conformer with reduced cgMLP and FFN intermediate dimensions to separate the architectural contribution from the additional parameters.

3.1 SPEECH-TO-TEXT

This setting follows Yano et al. (2026), who showed that depth up-scaling reaches ASR performance comparable to full fine-tuning with far less text degradation, with E-Branchformer expansion layers performing best. We revisit it with MA-Conformer expansion layers, which apply $\mathrm{cgMLP}_{\mathrm{1D}}$ to speech tokens (Section 2.2), and add SmolLM3-3B-Base and a parameter-matched comparison.

[2]This yields ranks much larger than typical LoRA configurations (e.g., $r = 16$), giving LoRA more capacity and making this a favorable setting for LoRA (Biderman et al., 2024). A rank sweep in Appendix F shows that smaller ranks reduce text degradation but increase WER.

Table 1: Speech-to-text results (INTERLEAVED). Depth up-scaling reaches WER comparable to full fine-tuning with far less text degradation, and MA-Conformer layers improve both. WER (%): LibriSpeech test-clean / test-other ↓. Text: avg. accuracy (%) over 8 benchmarks and Δ from pre-trained. *Italic* = full fine-tuning. **Bold** = best among depth up-scaling. †Reported by Tian et al. (2025). ‡From Yano et al. (2026) (1.7B only).

Model	Method	Trainable	WER ↓		Text ↑	
			clean	other	Avg.	Δ
Whisper-Medium† (769M)			2.8	6.5	—	—
1.7B	Pre-trained	—	—	—	69.0	0.0
	Full fine-tuning	1.87B	*2.3*	*5.6*	36.5	−32.5
	LoRA	0.66B	3.2	7.1	33.4	−35.6
	Depth Up-Scaling	0.66B	2.4	5.5	60.8	−8.2
	+ E-Branchformer‡	0.75B	**2.3**	5.3	**62.3**	**−6.7**
	+ MA-Conformer (matched)	0.66B	**2.3**	**5.1**	62.2	−6.8
	+ MA-Conformer	0.81B	**2.3**	**5.1**	62.1	−6.9
3B	Pre-trained	—	—	—	78.0	0.0
	Full fine-tuning	3.39B	*2.1*	*5.0*	39.1	−38.9
	LoRA	1.31B	2.9	6.2	40.0	−38.0
	Depth Up-Scaling	1.31B	2.3	5.1	58.8	−19.2
	+ MA-Conformer	1.59B	**2.0**	**4.7**	**60.2**	**−17.8**

3.1.1 EXPERIMENTAL SETUP

As in Yano et al. (2026), we train on the English ASR subset of the OWSM v3.2 suite (Tian et al., 2024) (approximately 48k hours) without mixing text data, to isolate the effect of depth up-scaling on text preservation[3]. We follow OpusLM (Tian et al., 2025) for speech tokenization and vocabulary expansion, as its tokenizer, models, and training data are fully open, its discrete tokens also support speech generation for future extensions, and its model sizes are close to ours. Full training configuration and training cost are in Appendices A and H.

We evaluate ASR using word error rate (WER) on LibriSpeech (Panayotov et al., 2015) test-clean and test-other with greedy decoding. For text capability, we report average accuracy across eight benchmarks (ARC-E/C (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2021), MMLU (Hendrycks et al., 2021)) with `olmes` (Gu et al., 2025). Since these benchmarks are all multiple-choice, we also evaluate SmolLM3-3B-Base on GSM8K (Cobbe et al., 2021), which requires generating a reasoning chain, and report the results in Appendix G. We evaluate with all added modules active. Since both LoRA and depth up-scaling can exactly recover the pre-trained model by removing the added parameters, this represents a stricter evaluation of text degradation.

3.1.2 RESULTS

Table 1 shows that depth up-scaling with standard transformer expansion layers, as in Yano et al. (2026), achieves WER comparable to full fine-tuning and clearly lower than LoRA, while preserving text far better than both, at 1.7B and 3B. Replacing them with MA-Conformer expansion layers further improves both ASR and text preservation at both scales. As a reference, depth up-scaling models at both scales match or surpass Whisper-Medium (Radford et al., 2023) on both test sets, a dedicated ASR model trained on 680k hours of supervised speech data. The parameter-matched MA-Conformer outperforms the transformer baseline on both ASR and text preservation and matches the full-size variant, indicating that the improvement cannot be explained by the additional parameters alone. The MA-Conformer also slightly improves test-other WER over E-Branchformer expansion layers (Kim et al., 2023), the best architecture in Yano et al. (2026), with comparable text preservation, even in its parameter-matched variant with fewer parameters.

[3]Depth up-scaling is orthogonal to text data replay, and the two can be combined.

Table 2: Image-to-text results (INTERLEAVED). Val. Loss: held-out validation loss ↓. Vision Avg.: mean accuracy (%) across ScienceQA, SEED-Bench, MMBench ↑. Text, italic, and bold as in Table 1. Conformer (1D) applies 1D convolution to image tokens instead of 2D.

Model	Method	Trainable	Val. Loss ↓	Vision Avg. ↑	Text ↑	
					Avg.	Δ
1.7B	Pre-trained	—	—	—	69.0	0.0
	Full fine-tuning	1.87B	*2.080*	*55.0*	66.6	−2.4
	LoRA	0.66B	2.349	49.8	68.0	−1.0
	Depth Up-Scaling (Transformer)	0.66B	2.225	53.0	**68.9**	**−0.1**
	+ MA-Conformer (matched)	0.66B	2.179	54.0	**68.9**	**−0.1**
	+ Conformer (1D)	0.81B	2.171	54.0	68.8	−0.2
	+ MA-Conformer	0.81B	**2.157**	**54.9**	**68.9**	**−0.1**
	+ Flamingo (gated cross-attn)	0.66B	2.245	50.4	68.6	−0.4
	+ Modality-Adaptive Attn	0.71B	2.244	51.6	68.4	−0.6
3B	Pre-trained	—	—	—	78.0	0.0
	Full fine-tuning	3.39B	*1.855*	*59.7*	72.4	−5.6
	LoRA	1.31B	2.158	54.1	76.7	−1.3
	Depth Up-Scaling (Transformer)	1.31B	1.948	58.0	75.8	−2.2
	+ MA-Conformer	1.59B	**1.934**	**58.7**	**76.3**	**−1.7**

3.2 IMAGE-TO-TEXT

We next apply the framework to image, a modality not covered by Yano et al. (2026). In the image instantiation (Section 2.3), image features from a frozen CLIP encoder are projected into the LLM, and expansion layers apply $\mathrm{cgMLP}_{\mathrm{2D}}$ (Section 2.2) to capture local spatial structure.

3.2.1 EXPERIMENTAL SETUP

We train on PixMo-Cap (Deitke et al., 2025) and COYO (Byeon et al., 2022), comprising approximately 1.4M image-text pairs in total, without instruction data or text-only data replay, following the pre-training stage approach in vision-language models (Lin et al., 2024; Shukor et al., 2025). Image tokens are produced by a frozen CLIP encoder (Radford et al., 2021) and projected into the LLM's hidden space via a learned linear layer. We compare along two axes on SmolLM2-1.7B. On the *method axis*, we compare against full fine-tuning and LoRA (Hu et al., 2022), and ablate four placement strategies for depth up-scaling (Appendix I). On the *architecture axis*, fixing placement to INTERLEAVED, we compare standard transformer, MA-Conformer, Flamingo-style gated cross-attention (Alayrac et al., 2022), and modality-adaptive attention (Ye et al., 2024) (separate key/value projections per modality in self-attention). At 3B, we compare full fine-tuning, LoRA, and transformer and MA-Conformer expansion layers. Full configuration is in Appendix B.

We evaluate CPT performance via validation loss on a held-out subset of the data that is never used for training. For downstream vision tasks, we report accuracy on ScienceQA (Lu et al., 2022), SEED-Bench (Li et al., 2024), and MMBench (Liu et al., 2024b), and use the average across the three benchmarks as the primary vision metric. Text capability is evaluated as in Section 3.1.

3.2.2 RESULTS

Table 2 reports results for both model scales. Depth up-scaling with transformer expansion layers approaches full fine-tuning in vision while using the same trainable parameter budget as LoRA, which underperforms by a clear margin. For text tasks, depth up-scaling causes negligible degradation across all conditions, possibly because the frozen CLIP encoder absorbs much of the modality gap and reduces the adaptation burden on the LLM layers. MA-Conformer expansion layers achieve the best validation loss and vision performance among depth up-scaling variants, with this advantage holding at both 1.7B and 3B. Among attention-based expansion layers, the standard transformer performs best, while Flamingo-style and modality-adaptive attention fall behind. In the image setting, the parameter-matched MA-Conformer still improves over the transformer baseline on both validation

Table 3: Jointly trained trimodal results (INTERLEAVED). The MA-Conformer gives the best ASR and vision performance among depth up-scaling variants while preserving text far better than full fine-tuning and LoRA. WER (%): LibriSpeech test-clean / test-other ↓. Vision Avg. and Text as in Tables 1 and 2. *Italic* and **bold** as in Table 1. [§]Separately trained unimodal MA-Conformer models from Tables 1 and 2 (Trainable: sum of the two).

Model	Method	Trainable	WER ↓		Vision Avg. ↑	Text ↑	
			clean	other		Avg.	Δ
1.7B	Pre-trained	—	—	—	—	69.0	0.0
	Full fine-tuning	1.87B	*2.5*	*5.4*	*47.3*	*35.8*	*−33.2*
	LoRA	0.66B	3.5	7.3	47.1	41.3	−27.7
	Depth Up-Scaling	0.66B	2.7	6.0	53.4	64.9	−4.1
	+ MA-Conformer (matched)	0.66B	**2.5**	5.9	54.0	**66.0**	**−3.0**
	+ MA-Conformer	0.96B	**2.5**	**5.8**	**54.4**	65.9	−3.1
	Unimodal MA-Conformer[§]	1.62B	2.3	5.1	54.9	—	—
3B	Pre-trained	—	—	—	—	78.0	0.0
	Full fine-tuning	3.39B	*2.4*	*5.5*	*47.7*	*38.9*	*−39.1*
	LoRA	1.31B	3.0	6.5	47.9	45.5	−32.5
	Depth Up-Scaling	1.31B	2.6	6.0	55.3	67.1	−10.9
	+ MA-Conformer	1.90B	**2.4**	**5.3**	**57.0**	**72.5**	**−5.5**
	Unimodal MA-Conformer[§]	3.18B	2.0	4.7	58.7	—	—

loss and vision accuracy, but by a smaller margin than in speech, suggesting that part of the full-size gain comes from the additional parameters. A Conformer that applies 1D convolution to image tokens at the same 0.81B falls between the transformer and the MA-Conformer, so matching the kernel to the 2D spatial structure of image tokens gives a further gain beyond adding a convolutional branch.

The above results are all CPT-stage evaluations. We also apply supervised fine-tuning (SFT) on the LLaVA-1.5 mix665k dataset (Liu et al., 2024a), after which the MA-Conformer still outperforms transformer expansion layers on all benchmarks (full results in Appendix J).

3.3 JOINTLY TRAINED TRIMODAL ADAPTATION

Finally, we adapt a single model to both speech and image, the main extension of this work, with both adaptations sharing the same expansion layers, and examine whether it approaches the unimodal results while preserving text ability. The MA-Conformer applies $\mathrm{cgMLP}_{\mathrm{1D}}$ to speech tokens and $\mathrm{cgMLP}_{\mathrm{2D}}$ to image tokens (Section 2.2), and we evaluate each modality separately.

3.3.1 EXPERIMENTAL SETUP

We train on a mixture of Localized Narratives (Pont-Tuset et al., 2020) (image captions recorded as spoken audio), PixMo-Cap (Deitke et al., 2025) combined with COYO (Byeon et al., 2022), and speech from the OWSM v3.2 ASR subset (Tian et al., 2024) and LibriSpeech (Panayotov et al., 2015), mixed at a 2:2:1 ratio (40%, 40%, and 20%, respectively). We train without text-only data replay, as in the unimodal settings. We use the same speech tokenizer and CLIP encoder as in Sections 3.1 and 3.2. We compare full fine-tuning, LoRA (Hu et al., 2022), and depth up-scaling with MA-Conformer expansion layers. In the trimodal setting, each expansion layer shares its MHSA (Equation (3)), FFN (Equation (5)), and all RMSNorm parameters across modalities, while maintaining separate cgMLP modules per modality (Equation (4)), including the gating projections and depthwise convolution, not only the convolution kernel. The parameter gap is larger in the trimodal setting due to the per-modality cgMLP duplication (0.96B vs. 0.66B at 1.7B), making the parameter-matched comparison most informative. Evaluation follows Sections 3.1 and 3.2, and full configuration is in Appendix C.

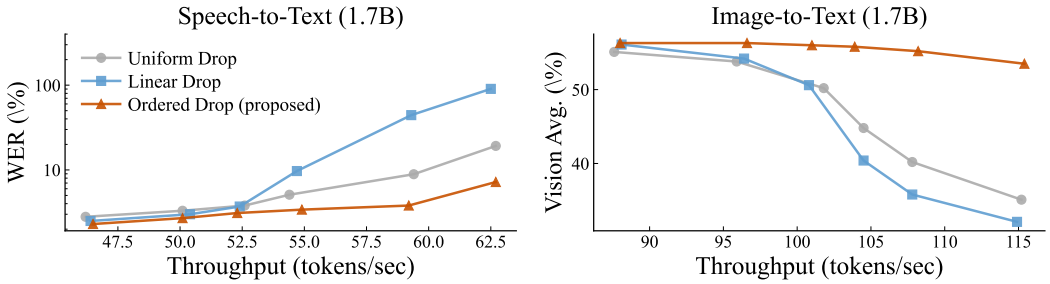

Figure 2: LibriSpeech test-clean WER ↓ (left, log scale) and Vision Avg. ↑ (right) vs. decoding throughput for SmolLM2-1.7B ($K = 6$) under three drop schedules, varying the number of active expansion layers k at inference ($k = K$ uses all expansion layers). Ordered Drop Pareto-dominates Uniform and Linear Drop across all k in both modalities.

3.3.2 Results

Table 3 shows that MA-Conformer expansion layers achieve the best ASR and vision performance among depth up-scaling variants while preserving text far better than full fine-tuning and LoRA, with this advantage holding at both 1.7B and 3B. Compared with the separately trained unimodal MA-Conformer models, the trimodal model shows only a small drop in both ASR and vision performance while using fewer trainable parameters than the two combined, and in a paired-input diagnostic on held-out Localized Narratives (Appendix E), providing the image slightly lowers WER. In the trimodal setting, the MA-Conformer also surpasses full fine-tuning on the vision average. LoRA achieves vision performance comparable to full fine-tuning but underperforms depth up-scaling on speech, which may reflect the limited capacity of low-rank updates for acoustic adaptation. At 1.7B in the trimodal setting, where the parameter gap is largest, the parameter-matched MA-Conformer again outperforms the transformer baseline and already performs close to the full-size variant.

3.4 Expansion LayerDrop

3.4.1 Experimental Setup

We evaluate Expansion LayerDrop in the two unimodal settings, which together cover both the 1D and 2D convolution paths of the MA-Conformer. Starting from the Interleaved MA-Conformer checkpoint of Section 3.1, we continue training on 3.8k hours of speech from the OWSM v3.2 ASR subset with each drop schedule of Section 2.4 applied only to the expansion layers, using the same optimizer settings. For Uniform Drop, $p = 0.5$. For Linear Drop, $p_{\min} = 0.05$ and $p_{\max} = 0.50$. We evaluate WER on LibriSpeech test-clean and decoding throughput (tokens/sec) for $k \in \{1, \ldots, K\}$ at inference. We apply the same procedure to the image checkpoint of Section 3.2, training for one additional epoch on the same image data, and evaluate Vision Avg. and throughput.

3.4.2 Results

Figure 2 plots task performance against decoding throughput as k varies from 1 to K. Ordered Drop Pareto-dominates Uniform and Linear Drop across all k values in speech, and the image setting shows the same trend. The other schedules collapse at small k, likely because the specific subsets used at inference are rarely seen during training, consistent with the benefit of aligning training and inference configurations. Appendix K shows consistent speech results on SmolLM3-3B-Base. Expansion LayerDrop thus enables a single trained model to serve multiple deployment scenarios. Selecting $k = K$ uses all expansion layers, reducing k trades performance for speed, and $k = 0$ recovers the original text LLM for text input.

4 Related Work

Multimodal LLMs. For speech, LLMs are extended either via encoder-based projection (Tang et al., 2024; Chu et al., 2024) or discrete speech tokens (Hassid et al., 2023; Nguyen et al., 2025;

Tian et al., 2025). For vision, models such as BLIP-2 (Li et al., 2023a) and LLaVA (Liu et al., 2023; 2024a) use a frozen image encoder (Radford et al., 2021) and a learned connector, keeping modality processing external to the LLM. A related line of work inserts modality-specific parameters directly into LLM layers. CogVLM (Wang et al., 2024) adds visual expert weights to attention and FFN, mPLUG-Owl2 (Ye et al., 2024) introduces separate key/value projections per token type, and Mixture-of-Transformers (Liang et al., 2025a) separates FFN, attention projections, and normalization by modality while keeping global self-attention. Our method follows this design but keeps the base LLM frozen and places modality-specific convolution only in the inserted expansion layers.

Forgetting in Multimodal CPT. CPT on new-modality data consistently degrades the text capabilities of the base LLM in both speech (Hsiao et al., 2025) and vision (Lin et al., 2024) settings. Text data replay (Tian et al., 2025; Peng et al., 2025) and parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) and Wings (Zhang et al., 2024b) are common mitigations, but replay requires access to the original pre-training data, and parameter-efficient methods constrain adaptation capacity. Alternatively, the LLM can be kept fully frozen while training only newly added modules that feed modality information into it, such as projections or cross-attention layers (Alayrac et al., 2022; Moon et al., 2024; Mo et al., 2025), preventing forgetting by design. Depth up-scaling likewise keeps the original weights frozen but inserts full expansion layers that process modality tokens together with text, enabling modality-adaptive processing at each depth.

Depth Up-Scaling. Depth up-scaling increases LLM capacity by inserting new transformer layers into a pre-trained model. SOLAR (Kim et al., 2024) introduces depthwise scaling for continued pre-training, and LLaMA Pro (Wu et al., 2024) shows that block expansion improves domain-specific adaptation while preserving general capability. Recent work has further improved initialization via learnable prediction (Yang et al., 2025b). These methods target text-domain expansion and build the inserted layers from copies of the base layers, whereas we design the inserted layers to follow the structure of each new modality. Yano et al. (2026) applied depth up-scaling to speech LLM adaptation with SmolLM2 models up to 1.7B, comparing expansion layer placements and transformer and E-Branchformer expansion layers. We extend it to image and jointly trained trimodal adaptation with expansion layers that apply 1D convolution to speech and 2D convolution to image tokens, scale it to 3B, and introduce Expansion LayerDrop.

5 Conclusion

We presented modality-adaptive depth up-scaling, which extends the depth up-scaling approach of Yano et al. (2026) from speech to a trimodal setting by inserting MA-Conformer expansion layers, which apply 1D convolution to speech tokens and 2D convolution to image tokens, into a frozen text LLM. Across speech-to-text, image-to-text, and jointly trained trimodal adaptation with 1.7B and 3B base LLMs, MA-Conformer expansion layers outperformed standard transformer expansion layers and caused less text degradation than full fine-tuning, and a single trimodal model approached the performance of separately trained unimodal models. We further introduced Expansion LayerDrop, which let a single trained speech or image model support multiple performance-speed operating points. These results indicate that depth up-scaling extends beyond a single new modality when the expansion layers follow the structure of each modality.

Limitations. Our findings are drawn from the settings we evaluated, namely two SmolLM base models at 1.7B and 3B parameters, English speech recognition, and vision-language evaluation mainly at the CPT stage. Expansion LayerDrop was evaluated on speech at both scales and on image at 1.7B, covering both convolution paths of the MA-Conformer, but not yet in the trimodal setting, where the shared expansion layers must support both modalities at each k. We leave this evaluation, including a per-modality choice of k, to future work. Because the framework only inserts layers into a standard decoder-only transformer and the trends held across base models of different generations and sizes, we expect them to carry over to other settings. Replication on other model families, larger scales, and multilingual speech is needed to confirm this. The parameter-matched comparisons are also limited to the 1.7B scale. When the expansion layers remain active, text degradation in the speech setting is far smaller than under full fine-tuning but not negligible, particularly at 3B and on generative reasoning. Beyond the paired-input diagnostic, our trimodal setting evaluates each modality separately and does not measure cross-modal reasoning over simultaneous speech and image input.

AI USE STATEMENT

We used generative AI tools at several stages of the study and manuscript preparation.

We consulted an LLM about implementation approaches for our method and used an LLM to review our experimental code. As the authors are non-native English speakers, we used an LLM to help draft English text from outlines and drafts written in our native language, and to edit the manuscript to improve readability. We also used generative AI tools to modify the plotting scripts for figures and to check the formatting of BibTeX entries.

Additionally, we used an LLM to check the internal consistency of the manuscript. This included feedback on how the experimental methodology was described and on whether the reported results supported the stated claims and interpretations. The authors assessed these suggestions and made all final scientific decisions.

We reviewed all AI-assisted work. We tested and reviewed the implementation to check that it behaved as intended, checked drafted and edited text for accuracy, checked AI-modified figures for consistency with the reported results, and verified all references against their original sources. The authors take responsibility for the final content of this work, including all text, claims, code, and figures produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

All base models (SmolLM2-1.7B, SmolLM3-3B-Base, CLIP ViT-L/14-336) and training datasets (OWSM v3.2, LibriSpeech, PixMo-Cap, COYO, Localized Narratives, LLaVA-1.5 mix665k) used in this work are publicly available. The depth up-scaling framework and the MA-Conformer are described in Section 2 and Appendix D, and expansion layer placements, optimizer settings, batch sizes, training durations, and tokenization details for each setting are given in Appendices A to C. Expansion LayerDrop schedules and their hyperparameters are described in Sections 2.4 and 3.4. Text benchmarks are evaluated with the publicly available `olmes` suite (Gu et al., 2025), vision benchmarks with `lmms-eval` (Zhang et al., 2024a), and speech benchmarks follow their standard evaluation protocols. All reported results are from a single training run per configuration. We will release our code upon publication.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in English and Mandarin. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/amodei16.html`.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro

von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro Von Werra, and Thomas Wolf. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7432–7439, 04 2020. doi: 10.1609/aaai.v34i05.6239.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ivCd8z8zR2. Featured Certification, Reproducibility Certification.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, June 2025.

Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. Stacking your transformers: A closer look at model growth for efficient LLM pre-training.

In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FXJDcriMYH.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.681. URL https://aclanthology.org/2024.acl-long.681/.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylO2yStDr.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=DgH9YCsqWm.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.282.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=UlHueVjAKr.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Donald Lane. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4fLr7H5D_eT.

Chi-Yuan Hsiao, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Wei-Chih Chen, and Hung-yi Lee. Analyzing Mitigation Strategies for Catastrophic Forgetting in End-to-End Training of Spoken Language Models. In *Interspeech 2025*, pp. 3234–3238, 2025. doi: 10.21437/Interspeech.2025-409.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe. E-Branchformer: Branchformer with enhanced merging for speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 84–91. IEEE, 2023.

Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 23–35, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.3. URL https://aclanthology.org/2024.naacl-industry.3/.

Sehoon Kim, Amir Gholami, Albert Eaton Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, and Kurt Keutzer. SqueezeFormer: An efficient transformer for automatic speech recognition. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=gE_vt-w4LhL.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27980–27993. Curran Associates, Inc., 2023. doi: 10.52202/075280-1214. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/58d0e78cf042af5876e12661087bea12-Paper-Conference.pdf.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, June 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

Yinan Liang, Ziwei Wang, Xiuwei Xu, Jie Zhou, and Jiwen Lu. EfficientLLaVA: Generalizable auto-pruning for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9445–9454, June 2025b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pp. 216–233, Berlin, Heidelberg, 2024b. Springer-Verlag. ISBN 978-3-031-72657-6. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=HjwK-Tc_Bc.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260.

Sicheng Mo, Thao Nguyen, Xun Huang, Siddharth Srinivasan Iyer, Yijun Li, Yuchen Liu, Abhishek Tandon, Eli Shechtman, Krishna Kumar Singh, Yong Jae Lee, Bolei Zhou, and Yuheng Li. X-fusion: Introducing new modality to frozen large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 228–238, October 2025.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. AnyMAL: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1314–1332, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.98.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pp. 2613–2617, 2019. doi: 10.21437/Interspeech.2019-2680.

Yifan Peng, Krishna C Puvvada, Zhehuai Chen, Piotr Zelasko, He Huang, Kunal Dhawan, Ke Hu, Shinji Watanabe, Jagadeesh Balam, and Boris Ginsburg. VoiceTextBlender: Augmenting large language models with speech capabilities via single-stage joint speech-text supervised fine-tuning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5787–5802, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.298.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pp. 647–664. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–23, 2025.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Jinchuan Tian, Yifan Peng, William Chen, Kwanghee Choi, Karen Livescu, and Shinji Watanabe. On the Effects of Heterogeneous Data Sources on Speech-to-Text Foundation Models. In *Interspeech 2024*, pp. 3959–3963, 2024. doi: 10.21437/Interspeech.2024-1938.

Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, Huck Yang, and Shinji Watanabe. OpusLM: A Family of Open Unified Speech Language Models. In *Interspeech 2025*, pp. 3259–3263, 2025. doi: 10.21437/Interspeech.2025-1184.

Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, 2024.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. LLaMA Pro: Progressive LLaMA with block expansion. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6518–6537, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.352. URL https://aclanthology.org/2024.acl-long.352/.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22–31, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yifei Yang, Zouying Cao, Xinbei Ma, Yao Yao, Zhi Chen, Libo Qin, and Hai Zhao. LESA: Learnable LLM layer scaling-up. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22463–22476, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1095. URL https://aclanthology.org/2025.acl-long.1095/.

Kazuki Yano, Jun Suzuki, and Shinji Watanabe. Adapting Text LLMs to Speech via Multimodal Depth Up-Scaling. In *Interspeech 2026*, pp. 6410–6415, 2026. doi: 10.21437/Interspeech.2026-2099.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 13040–13051, 2024.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24972–24982, June 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Wings: Learning multimodal LLMs without text-only forgetting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=nqWaya7hiX.

A Speech-to-Text Training Configuration

All speech-to-text experiments are run on a single node of 4 NVIDIA H100 GPUs.

Optimizer. We use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and weight decay 0.1. The learning rate follows a linear warmup from 0 to 1×10^{-4} over the first 10,000 steps, then decays linearly to zero over 125,000 total steps. Gradients are clipped to a maximum ℓ_2-norm of 1.0. All models are trained in bfloat16 precision.

Batch size and steps. We use token-count batching with 10,000 tokens per GPU and a micro-batch size of 1, with a maximum context length of 8,192 tokens. Models are trained for up to 125,000 steps, with checkpoints saved every 5,000 steps, and we report results from the final checkpoint. Training data is the English ASR subset of OWSM v3.2 (Tian et al., 2024), comprising approximately 48k hours, as described in Section 3.1.

Speech tokenization. Discrete speech tokens are produced by a two-stage pipeline following OpusLM (Tian et al., 2025). Continuous HuBERT (Hsu et al., 2021) features are extracted from layer 18 of an ESPnet HuBERT checkpoint trained for 38 epochs, then quantized via k-means with $k = 5{,}000$ cluster centroids, yielding 5,000 discrete self-supervised learning (SSL) tokens. Audio is additionally tokenized with a DAC-based neural codec (Kumar et al., 2023) using 8 codebooks per frame, yielding 8,192 codec tokens, so that each frame is represented by one SSL token and 8 codec tokens. The full vocabulary contains 62,600 tokens: 256 special tokens, 5,000 SSL tokens, 8,192 codec tokens, and 49,152 text byte-pair encoding (BPE) tokens shared with the base LLM.

SpecAugment. For ASR sequences, we apply time masking (Park et al., 2019) with 10 masks, each spanning up to 5% of the sequence length.

Expansion layer placement. One MA-Conformer expansion block is inserted after every 4th frozen transformer layer (INTERLEAVED placement, Section 2.1), that is, after original layers $\{4, 8, 12, 16, 20, 24\}$ for SmolLM2-1.7B ($K = 6$) and $\{4, 8, \ldots, 36\}$ for SmolLM3-3B-Base ($K = 9$), corresponding to $\mathcal{S} = \{5, 10, \ldots, 30\}$ and $\{5, 10, \ldots, 45\}$ in Equation (1). Each expansion block uses a depthwise convolution kernel size of 31 and a pointwise feed-forward projection whose hidden size matches the base LLM's MLP hidden size (8,192 for SmolLM2-1.7B and 11,008 for SmolLM3-3B-Base). Following prior work (Wu et al., 2024), the MHSA, FFN, and RMSNorm weights of each expansion layer are initialized by copying the adjacent lower frozen layer. Following Yano et al. (2026), the output projection of the cgMLP is initialized to zero, so the cgMLP initially adds nothing to the residual stream. In the parameter-matched variant, whose FFN and cgMLP dimensions differ from those of the base layers, the FFN and cgMLP are instead initialized randomly. For depth up-scaling, all original LLM layers and the embedding and output rows of the original text vocabulary are frozen throughout training and excluded from weight decay, so they remain identical to those of the base LLM. Only the inserted MA-Conformer weights and the embedding and output rows of the added vocabulary (special, SSL, and codec tokens) are updated. In all experiments, modality spans precede the text tokens and the training loss is computed only on text tokens, so the non-causal convolutions (Appendix D) do not expose future prediction targets. For text-only deployment, removing the expansion layers together with these added rows exactly recovers the base LLM. In all tables, Trainable counts the parameters of all trained modules together with the full embedding and output matrices (0.26B at 1.7B and 0.58B at 3B), using the same convention for every method. For depth up-scaling and LoRA, only the rows of the added vocabulary in these matrices are updated (none in the image setting), so the reported counts are upper bounds on the parameters actually updated. Excluding these matrices, depth up-scaling with transformer expansion layers and LoRA both train about 0.40B parameters at 1.7B.

B IMAGE-TO-TEXT TRAINING CONFIGURATION

All image-to-text experiments are run on a single node of 4 NVIDIA H100 GPUs.

Optimizer. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.1 and gradient clipping to a maximum ℓ_2-norm of 1.0. The learning rate follows a linear warmup from 0 to 1×10^{-5} over the first 10% of training steps, then decays linearly to zero. All models are trained in bfloat16 precision with a maximum sequence length of 2,048 tokens.

Batch size and steps. We use a per-device batch size of 4 with no gradient accumulation. Models are trained for 1 epoch over the training data (PixMo-Cap (Deitke et al., 2025) and COYO (Byeon et al., 2022), approximately 1.4M image-text pairs, as described in Section 3.2).

Image encoding. Image features are extracted from the penultimate layer of a frozen CLIP ViT-L/14-336 encoder (Radford et al., 2021), producing 576 (24×24) patch tokens per image. These are

Table 4: Data mixture for trimodal training. Sampling weights are relative; the effective ratio groups all speech sources.

Dataset	Modality	Weight	Eff. ratio
Localized Narratives (Pont-Tuset et al., 2020)	Image+speech → text	3.0	2
PixMo-Cap (Deitke et al., 2025) + COYO (Byeon et al., 2022)	Image → text	3.0	2
OWSM v3.2 (Tian et al., 2024)	Speech → text	0.5	1
LibriSpeech (Panayotov et al., 2015)	Speech → text	1.0	

projected into the LLM's hidden space via a trainable linear projection. For depth up-scaling, the CLIP encoder and all original LLM weights are frozen throughout training, and only the expansion layers and the linear projector are updated. Full fine-tuning also keeps the CLIP encoder frozen.

Expansion layer placement. The interleaved placement follows the same scheme as the speech experiments (Appendix A). Each MA-Conformer expansion block applies $\mathrm{cgMLP}_{\mathrm{2D}}$ with a 3×3 depthwise convolution and an intermediate feed-forward size matching the base LLM's MLP hidden size (8,192 for SmolLM2-1.7B and 11,008 for SmolLM3-3B-Base), inserted after the self-attention module. Initialization follows Appendix A.

Evaluation. ScienceQA, SEED-Bench, and MMBench are evaluated with `lmms-eval` (Zhang et al., 2024a).

C Trimodal Training Configuration

All trimodal experiments are run on a single node of 4 NVIDIA H100 GPUs.

Optimizer. Optimizer settings follow Appendix A: AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, weight decay 0.1, and gradient clipping to a maximum ℓ_2-norm of 1.0. The learning rate follows a linear warmup from 0 to 1×10^{-4} over the first 10,000 steps, then decays linearly to zero over 250,000 total steps. All models are trained in bfloat16 precision.

Batch size and steps. We use token-count batching with 32,000 tokens per GPU and a micro-batch size of 1. Models are trained for up to 250,000 steps, with checkpoints saved every 5,000 steps.

Modality encoders. Speech tokenization follows Appendix A. Image encoding follows Appendix B: features from the penultimate layer of a frozen CLIP ViT-L/14-336 (Radford et al., 2021) projected via a trainable linear layer.

Expansion layer placement. Expansion layers are placed identically to the unimodal experiments (Appendix A). Each MA-Conformer block applies $\mathrm{cgMLP}_{\mathrm{1D}}$ to speech tokens and $\mathrm{cgMLP}_{\mathrm{2D}}$ to image tokens as described in Section 2.2. Initialization follows Appendix A.

Data mixture. Table 4 shows the four data sources and their sampling weights. The effective ratio is 2:2:1 (image-narrative : image-caption : speech), as described in Section 3.3. In all experiments, each modality forms a single contiguous span that precedes the text tokens. When both modalities are present, the image span precedes the speech span. For LibriSpeech, we use only the 960-hour training split, and test-clean and test-other are excluded from all training and model selection.

D MA-Conformer Details

Each cgMLP follows the convolutional gating MLP of Kim et al. (2023). Given input $\mathbf{X} \in \mathbb{R}^{n \times d}$, it computes

$$[\mathbf{Z}_1; \mathbf{Z}_2] = \mathrm{SiLU}(\mathbf{X}\mathbf{W}_1), \tag{6}$$

$$\mathrm{cgMLP}(\mathbf{X}) = \big(\mathbf{Z}_1 \odot \mathrm{DWConv}(\mathrm{LN}(\mathbf{Z}_2))\big)\mathbf{W}_2, \tag{7}$$

Table 5: WER (%) ↓ on 1k held-out Localized Narratives examples for the jointly trained SmolLM2-1.7B MA-Conformer model.

Input	WER
Speech only	10.0
Speech + image	**9.7**

Table 6: LoRA rank sweep for speech-to-text on SmolLM2-1.7B. WER (%): LibriSpeech test-clean / test-other ↓. Text Δ: change in average text accuracy from pre-trained. Trainable counts include the embedding and output layers.

Method	Trainable	clean	other	Text Δ
LoRA ($r = 16$)	0.27B	6.2	15.8	-10.8
LoRA ($r = 64$)	0.33B	5.3	14.2	-17.8
LoRA ($r = 356$)	0.66B	3.2	7.1	-35.6
MA-Conformer (matched)	0.66B	**2.3**	**5.1**	$\mathbf{-6.8}$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{\mathrm{cg}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{\mathrm{cg}}/2 \times d}$, $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{n \times d_{\mathrm{cg}}/2}$ split the projection along channels, LN is layer normalization, and DWConv is a depthwise convolution. $\mathrm{cgMLP}_{\mathrm{1D}}$ uses a 1D kernel of size 31 over the speech sequence, and $\mathrm{cgMLP}_{\mathrm{2D}}$ reshapes the image tokens into their 24×24 patch grid and uses a 3×3 kernel. Both convolutions are non-causal and use zero padding that preserves the input size (15 for the 1D kernel and 1 for the 2D kernel). The image tokens are the 576 CLIP patch tokens, excluding the CLS token, arranged on the grid in row-major order. Each input contains a single image in all our experiments. The intermediate size d_{cg} equals the MLP hidden size of the base LLM (8,192 for SmolLM2-1.7B and 11,008 for SmolLM3-3B-Base). In the parameter-matched variant, the cgMLP and FFN intermediate sizes are reduced together so that the trainable parameters match those of the transformer baseline (0.66B at 1.7B). In the unimodal settings, both are reduced from 8,192 to 5,376, and in the trimodal setting, where each layer has two cgMLP modules, they are reduced further to meet the same budget.

E Paired Speech-and-Image Diagnostic

To characterize how the jointly trained model handles paired inputs, we evaluate the SmolLM2-1.7B MA-Conformer model of Section 3.3 on 1k Localized Narratives examples held out from training, each pairing an image with a spoken description of it. We measure WER on the spoken description with greedy decoding, with and without the image in the input. As shown in Table 5, providing the image gives a small improvement, from 10.0 to 9.7 WER.

F LoRA Configuration and Rank Sweep

We apply LoRA (Hu et al., 2022) to all attention (query, key, value, output) and MLP (gate, up, down) projections in every layer, with $\alpha = r$ and no dropout. All other settings, including the learning rate schedule and the number of training steps, follow the depth up-scaling configuration in Appendix A. In the speech experiments, r is set to match the trainable parameter count of depth up-scaling, giving $r = 356$ for SmolLM2-1.7B and $r = 384$ for SmolLM3-3B-Base.

Table 6 shows a rank sweep on SmolLM2-1.7B. Lower ranks reduce text degradation but increase WER, so the rank moves LoRA along a trade-off between ASR and text preservation. None of the evaluated ranks reaches the parameter-matched MA-Conformer, which is better on both.

G Generative Reasoning on GSM8K

We evaluate GSM8K (Cobbe et al., 2021) on SmolLM3-3B-Base with all expansion layers active, as for the other text benchmarks. Following the SmolLM3 evaluation (Bakouch et al., 2025), we use `olmes` with 5-shot prompting. We focus on SmolLM3-3B-Base, whose pre-trained GSM8K

Table 7: GSM8K accuracy (%) ↑ for SmolLM3-3B-Base with all expansion layers active.

Method	Speech	Image
Pre-trained	67.1	67.1
Full fine-tuning	1.2	54.4
Depth Up-Scaling + MA-Conformer	**16.1**	**66.9**

Table 8: Training cost for speech-to-text on SmolLM2-1.7B (4 NVIDIA H100 GPUs).

Method	Trainable	GPU-hours
Full fine-tuning	1.87B	116
LoRA ($r = 356$)	0.66B	131
Depth Up-Scaling	0.66B	97
+ MA-Conformer	0.81B	124

accuracy (67.1) is much higher than the reported 5-shot accuracy of SmolLM2-1.7B (31.0; Ben Allal et al., 2025), which makes degradation easier to measure. Table 7 shows that in the image setting, depth up-scaling keeps GSM8K accuracy nearly unchanged, whereas full fine-tuning degrades it. In the speech setting, all methods degrade more, and depth up-scaling retains part of the capability while full fine-tuning loses it almost entirely.

H Training Cost

Table 8 reports the training cost in the speech setting on SmolLM2-1.7B. All methods run the same number of steps on the same hardware with the same implementation, and every step runs the full base model. GPU-hours therefore fall within about 1.4× of each other, since the forward and backward passes through the base model dominate the cost. Depth up-scaling is the fastest because the frozen base layers need no weight gradients. The MA-Conformer adds the cost of the convolutional branch, and LoRA at $r = 356$ adds low-rank matrix multiplications to every linear layer, making it the slowest. Our method targets preserving the text capabilities of the base LLM rather than reducing training cost, and its cost remains comparable to that of full fine-tuning.

I Image-to-Text Placement Ablation

Table 9 shows the full placement ablation for image-to-text on SmolLM2-1.7B, with individual benchmark scores. As in prior work on speech (Yano et al., 2026), INTERLEAVED achieves the best task performance.

J Image-to-Text SFT Results

To examine whether the CPT-stage trends in Table 2 carry over after instruction tuning, we apply SFT to the SmolLM2-1.7B CPT checkpoints on the LLaVA-1.5 mix665k dataset (Liu et al., 2024a).

SFT configuration. We fine-tune each CPT checkpoint for one epoch (20,791 steps) with AdamW (Loshchilov & Hutter, 2019), using a peak learning rate of 1×10^{-5}, a 1,000-step linear warmup, and linear decay to zero. The CLIP encoder remains frozen. Full fine-tuning updates all LLM parameters, LoRA updates the adapters and the projector, and depth up-scaling updates the expansion layers and the projector. Other settings follow Appendix B.

Results. Table 10 shows that the MA-Conformer outperforms transformer expansion layers on all benchmarks after SFT and matches or exceeds full fine-tuning on three of the four. At 1.7B, it approaches LLaVA-1.5 (Liu et al., 2024a) on POPE and surpasses it on ScienceQA, despite the latter using a 7B backbone.

Table 9: Image-to-text placement ablation for SmolLM2-1.7B. Val. Loss: held-out validation loss ↓. Vision: accuracy (%) on ScienceQA (Lu et al., 2022), SEED-Bench (Li et al., 2024), MMBench (Liu et al., 2024b); Avg.: mean across the three ↑. Text: avg. accuracy (%) across 8 benchmarks; Δ from pre-trained. *Italic* = full fine-tuning. **Bold** = best among depth up-scaling.

Method	Trainable	Val. Loss ↓	Vision ↑				Text ↑	
			ScienceQA	SEED	MMBench	Avg.	Avg.	Δ
Pre-trained	—	—	—	—	—	—	69.0	0.0
Full fine-tuning	1.87B	*2.080*	*59.9*	*47.1*	*58.1*	*55.0*	66.6	−2.4
Depth Up-Scaling (BOTTOM)	0.66B	2.302	57.2	39.7	49.9	48.9	68.7	−0.3
Depth Up-Scaling (TOP)	0.66B	2.339	56.5	38.2	44.0	46.2	**69.0**	**0.0**
Depth Up-Scaling (SANDWICH)	0.66B	2.240	58.1	43.0	55.0	52.0	68.9	−0.1
Depth Up-Scaling (INTERLEAVED)	0.66B	**2.225**	**59.2**	**44.5**	**55.2**	**53.0**	68.9	−0.1

Table 10: Image-to-text SFT results for SmolLM2-1.7B. POPE: accuracy (%) on the random split (Li et al., 2023b) ↑. MME: perception score (Fu et al., 2025) ↑. ScienceQA (Lu et al., 2022), SEED-Bench (Li et al., 2024): accuracy (%) ↑. *Italic* = full fine-tuning. **Bold** = best among depth up-scaling. †Results from the original paper; model uses a 7B LLM backbone.

Method	POPE ↑	MME ↑	ScienceQA ↑	SEED ↑
LLaVA-1.5† (Liu et al., 2024a)	87.3	1510.7	66.8	66.1
Full fine-tuning	*85.3*	*1312.7*	*66.1*	*62.3*
LoRA ($r = 356$)	83.9	1110.8	66.6	54.3
Depth Up-Scaling	82.8	1322.4	67.8	62.7
+ MA-Conformer	**85.1**	**1361.2**	**68.6**	**62.8**

K ADDITIONAL EXPANSION LAYERDROP RESULTS

Figure 3 shows Expansion LayerDrop results for speech on SmolLM3-3B-Base ($K = 9$), trained with the same procedure as Section 3.4. For all Expansion LayerDrop runs, the optimizer state and learning rate schedule, including warmup, are restarted from the adapted checkpoint. As at 1.7B, Ordered Drop Pareto-dominates Uniform and Linear Drop at all k except $k = 6$, where Linear Drop is marginally better, while the other schedules collapse at small k. In Figures 2 and 3, each Uniform Drop point is a single evaluation run in which the layer subset is drawn at random for each input, whereas Linear Drop deterministically uses the first k expansion layers, which have the lowest drop probabilities. For a given k, all schedules run the same number of expansion layers, so their throughput is identical in principle, and the small differences in Figures 2 and 3 reflect measurement noise.

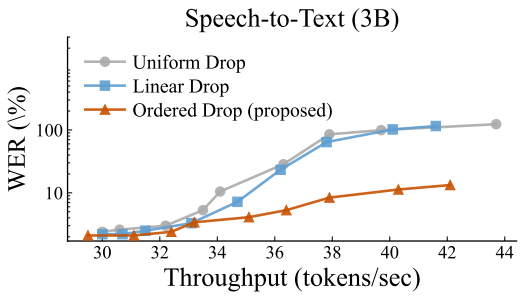

Figure 3: LibriSpeech test-clean WER ↓ (log scale) vs. decoding throughput for SmolLM3-3B-Base ($K = 9$) under three drop schedules, varying the number of active expansion layers k at inference.