

LF²AR: Accounting for Layerwise Dynamics to Improve Multimodal Adaptation of Language Models

Santiago Cuervo¹ Adel Moumen² Yanis Labrak⁶ Sameer Khurana^{3,†}
 Antoine Laurent^{5,†} Mickael Rouvier⁴ Phil Woodland² Ricard Marxer^{1,7}

¹Université de Toulon, Aix-Marseille Université, CNRS, LIS; ²Department of Engineering, University of Cambridge; ³Mitsubishi Electric Research Laboratories; ⁴LIA, Avignon Université; ⁵LIUM, Le Mans Université; ⁶Idiap Research Institute; ⁷CNRS, ILLS

[†]Contributed while at the listed affiliations.

Abstract

Text-pretrained language models (LMs) encode rich world knowledge, but adapting them to process and generate perceptual modalities such as audio and images while effectively leveraging that knowledge remains challenging. Perceptual modalities are finer-grained and less semantically dense than text, making it unclear how functions learned during text pretraining can be reused. We study this problem through the lens of a layerwise abstraction-refinement dynamic observed in transformer LMs: representations first become more abstract and compositional, then are refined into representations predictive of fine-grained structure. This perspective suggests that adapting an LM to finer-grained modalities requires: (i) allocating additional fine-to-coarse processing at the input and coarse-to-fine processing at the output, consistent with late modality fusion and an output-side analogue we term late fission; and (ii) allowing the output predictor to preserve input-dependent selective access to both high-level semantic structure and low-level perceptual detail, motivating our use of attention residuals in fission. We instantiate this view in LF²AR, a simple architecture combining these mechanisms, and study it on text-as-images and speech, two modalities for which semantic correspondence to text can be controlled. Across models ranging from 135M to 2B parameters and adapted to these modalities, we find that these components increase feature abstraction, strengthen cross-modal alignment, enable such alignment to emerge at smaller compute budgets, improve preservation of text-like predictive structure, and yield better performance on text-as-images and speech versions of language understanding and reasoning benchmarks. Additionally, attention residuals induce sparse, interpretable use of deep backbone layers, enabling early-exit decoding with a $1.9\times$ generation speedup.

1 Introduction

The remarkable capabilities of large language models (LLMs) trained on web-scale text corpora, together with the relative scarcity of equally rich semantic supervision in other modalities, have made transfer of text-pretrained knowledge a central goal in multimodal AI (Alayrac et al., 2022; Li et al., 2023; Girdhar et al., 2023; Shi et al., 2025). A common strategy is to fine-tune pretrained text LLMs on multimodal corpora and instruction-tuning data so that they can consume and generate non-textual signals while reusing capabilities acquired during text pretraining. This recipe underlies much recent progress in multimodal language modeling. Yet the resulting transfer is often imperfect: models frequently underperform when semantically equivalent content is presented in non-textual rather than textual form (Tang et al., 2025; Venhoff et al., 2025; Cuervo et al., 2026), and multimodal adaptation can also induce forgetting of text-only capabilities (Zhang et al., 2024; Cuervo et al., 2026). These failures highlight the challenge of adapting LLMs to multimodal signals in a way that preserves and reuses what was learned during text pretraining.

We study the challenge of multimodal adaptation primarily as one of cross-modal transfer: semantically equivalent content should engage similar computations across modalities. For example, the sentence “*Your submission has been accepted*” should light up similar neural circuits of positive emotions, whether heard aloud or read in an email. A central obstacle to such transfer is a mismatch in representational granularity. Text token sequences are semantically dense, with each token typically corresponding to a word-like unit. By contrast, visual or audio signals are represented as much longer streams of lower-level units—pixels or patches and audio frames, respectively—over which semantics are distributed more sparsely along spatial or temporal dimensions and mixed with modality-specific information. As a result, cross-modal transfer is non-trivial: even when the underlying semantics are the same, the representation space of perceptual modalities differs substantially from the input space over which text-learned functions are defined.

A large body of prior work has attempted to address these challenges through complex multi-stage training pipelines and modular architectures with adapters that bridge the granularity mismatch between perceptual signals and language representations, both at the input—a family of methods often termed as doing *late modality fusion* (Alayrac et al., 2022; Li et al., 2023; Jiang et al., 2024; Yang et al., 2024; Shi et al., 2025; Tseng et al., 2026)—and at the output (Li et al., 2023; Dong et al., 2024; Xu et al., 2025), which we analogously term *late modality fission*¹. Many other works, however, adopt and advocate for simpler, more monolithic architectures that rely on lightweight linear mappings for multimodal adaptation (Chameleon Team, 2024; Nguyen et al., 2025; Zeng et al., 2025; Wu et al., 2025; Cui et al., 2025). These methods are commonly referred to as *native* multimodal LMs or, in our preferred terminology, as performing *early modality fusion*; correspondingly, we refer to their output-side analogue as *early modality fission*. The trade-offs between these approaches remain an active area of research (Shukor et al., 2025; Venhoff et al., 2025; Diao et al., 2026).

In this work, rather than starting from complex multimodal adaptation pipelines or from the early-versus-late fusion/fission dichotomy alone, we ask what the internal organization of a text-pretrained LM implies about how multimodal adaptation should be structured. Our starting point is recent work suggesting that transformer LMs exhibit a consistent two-phase representational dynamic across layers: earlier layers tend to compose lower-level features into more abstract and compositional ones, whereas later layers refine these abstractions into lower-level features predictive of upcoming finer-grained structure (Valeriani et al., 2023; Antonello & Cheng, 2024; Cheng et al., 2025; Queipo-de-Llano et al., 2026). We refer to this as the *layerwise abstraction-refinement dynamics*. Because perceptual signals are finer-grained than text, effective adaptation that preserves and leverages this dynamic should provide additional modality-specific layers for fine-to-coarse processing before the pretrained backbone and coarse-to-fine processing after it. We study these as simple instances of *late fusion* and *late fission*, respectively, and analyze their effects on representation dynamics across layers to better understand their relation to cross-modal transfer.

Additionally, we hypothesize that, given the finer granularity of perceptual signals in time and/or space relative to text, generating them should rely only sparsely on higher-level semantic features and more often on low-level local detail. For example, in the context of speech generation, given the utterance “‘we submitted the paper to the ’”, predicting upcoming speech tokens may initially benefit from high-level features, such as anticipating “‘conference’”. However, once the model begins generating “‘conference’”, accurate prediction depends on fine-grained local context—for instance, knowing that the current phone is ‘f’ is necessary to predict the next phone ‘e’. This leads us to propose a novel design primitive based on *attention residuals* (similar to Kimi Team et al. (2026)), which gives the fission mechanism dynamic access to multiple representation levels and allows it to switch between coarse semantic information and fine perceptual detail as needed.

Our main contributions can be summarized as follows:

¹Perhaps a non-standard term, but with precedents in the literature (Landragin, 2007; Marinov et al., 2023).

- We introduce simple architectural instantiations of late fusion and late fission, together with a framework for studying their effects on cross-modal transfer through the lens of layerwise abstraction-refinement dynamics.
- We propose an additional design primitive for multimodal adaptation of LMs: attention residuals to augment modality fission.
- We apply this framework to the study of text LMs adapted for generation of text-as-images and speech, and through extensive ablations demonstrate the effects of these multimodal adaptation architectural choices.
- We show that fission with attention residuals learns sparse use of backbone depth, and leverage it to speed up generation through dynamic compute allocation.

2 Preliminaries

Multimodal adaptation of language models. Let $\mathbf{w} = (w_1, \dots, w_T)$ denote a sequence of text tokens, with $w_i \in \mathcal{V}_{\text{text}}$. A text LM P_θ defines the autoregressive distribution $P_\theta(\mathbf{w}) = \prod_{i=1}^T P_\theta(w_i \mid \mathbf{w}_{<i})$, and is pretrained by maximum likelihood over a text corpus, $\mathcal{D}_{\text{text}}$: $\mathcal{L}_{\text{text}}(\theta) = -\sum_{\mathbf{w} \in \mathcal{D}_{\text{text}}} \sum_{i=1}^{|\mathbf{w}|} \log P_\theta(w_i \mid \mathbf{w}_{<i})$. We study *multimodal adaptation* as a second stage of training applied to such a pretrained text LM. Our focus is on models adapted not only for multimodal understanding, but also for multimodal generation, through continued autoregressive modeling of sequences interleaving text with tokens from an additional modality. Let $\mathbf{x} = (x_1, \dots, x_T)$ denote such a sequence, where each token belongs either to the text vocabulary $\mathcal{V}_{\text{text}}$ or to the vocabulary $\mathcal{V}_m$ of one target modality m , i.e., $x_i \in \mathcal{V}_{\text{mm}} = \mathcal{V}_{\text{text}} \cup \mathcal{V}_m$. Starting from the pretrained LM, multimodal adaptation yields a model $P_{\theta'}$, where θ' may include both the original backbone parameters and any additional modality-specific parameters. $P_{\theta'}$ is trained to autoregressively model $\mathbf{x}$:

$$\mathcal{L}_{\text{mm}}(\theta') = - \sum_{\mathbf{x} \in \mathcal{D}_{\text{mm}}} \sum_{i=1}^{|\mathbf{x}|} \log P_{\theta'}(x_i \mid \mathbf{x}_{<i}). \quad (1)$$

In this paper, we study two concrete instances of this setup: LMs adapted for *pixel language modeling* (Rust et al., 2023) and *speech language modeling* (Lakhota et al., 2021). In both cases, multimodal sequences are constructed by switching between text and the target modality at the word level according to a fixed interleaving rate, similarly to Nguyen et al. (2025). In the pixel LM setting, the additional modality consists of rendered text images represented as sequences of flattened binary image patches $\mathbf{p}_j \in \{0,1\}^K$. The adapted model predicts each patch through a factorized Bernoulli distribution over its pixels, $P_{\theta'}(\mathbf{p}_j \mid \mathbf{x}_{<j}) = \prod_{k=1}^K P_{\theta'}(p_{j,k} \mid \mathbf{x}_{<j})$, conditioned on the preceding multimodal context. In the speech LM setting, the additional modality consists of vector-quantized self-supervised speech representations, as in Lakhota et al. (2021). These speech tokens are treated as elements of a modality-specific vocabulary $\mathcal{V}_{\text{speech}}$.

Abstraction-refinement dynamics in transformer language models. Recent work suggests a consistent picture of how transformer LMs organize computation across depth. Rather than changing monotonically, hidden representations appear to pass through two broad phases, which we refer to jointly as the *layerwise abstraction-refinement dynamic*. In the *abstraction* phase, lower-level features are progressively composed into coarser and more abstract ones; geometric analyses based on intrinsic dimensionality and neighborhood structure associate this phase with an expansion of the representation manifold, suggesting that the model is constructing a richer feature basis in which linguistic structure can be represented (Valeriani et al., 2023; Cheng et al., 2025). A useful intuition comes from the compositional nature of language: in a character-level LM, for example, a vocabulary of a few dozen symbols must be composed into a far larger space of morphemes, words, and multi-word meanings. In the subsequent *refinement* phase, the model increasingly shifts toward deriving lower-level features conditioned on the coarser representations composed earlier, in ways that are useful for next token prediction (Cheng et al., 2025; Antonello & Cheng, 2024; Queipo-de-Llano et al., 2026).

3 Accounting for Layerwise Abstraction-Refinement Dynamics

Architectural hypotheses. A lesson from the abstraction-refinement literature is that different layers tend to operate at characteristic abstraction levels. Consequently, text-pretrained computations are not uniformly reusable across depth: a circuit trained to respond to word-level features, for example, may fail to activate if the representations reaching it are characters. Because perceptual modalities are finer-grained than text, this view suggests the need for additional processing on both sides of the pretrained backbone.

Hyp. Multimodal adaptation benefits from late fusion for additional fine-to-coarse processing before the pretrained text backbone and late fission for additional coarse-to-fine processing after it to better match expected layerwise abstraction levels.

The abstraction-refinement literature also suggests that such additional composition and refinement need not rely on the specialized training algorithms and architectures often used in late fusion/fission. The standard language modeling objective with transformer layers already appears sufficient to induce these phases in text-pretrained backbones. We therefore adopt such a minimalist approach, implementing late fusion and late fission as additional layers placed before and after the backbone, respectively, applied only to the target modality, and trained end-to-end with the same autoregressive objective as the rest of the model.

Regarding how fine-grained modeling should use the pretrained model, we posit that high-level text features may be most useful when selecting the next linguistic unit, whereas low-level local structure may matter more when unfolding that unit into a sequence of perceptual features. If so, effective adaptation should avoid forcing such low-level detail to be propagated throughout the pretrained backbone, where it may interfere with the formation and reuse of higher-level text-like features. For example, as illustrated in the introduction, in speech generation, predicting the beginning of a word may benefit from text-like features that anticipate the next lexical item, while predicting the following phones within that word depends on local phonetic context.

Hyp. Fine-grained modeling should rely on both higher-level guidance and lower-level detail as needed; thus fission benefits from dynamic access to multiple representation levels, rather than propagating low-level detail through the pretrained backbone.

We instantiate this hypothesis through *attention residuals*. Let $\mathbf{c}_i^{(1)}, \dots, \mathbf{c}_i^{(L)} \in \mathbb{R}^d$ denote the residual-stream² representations at position i from the L layers of the model up to the backbone output. In a standard residual operation, a layer augments its current representation with a fixed skip contribution from an earlier computation. Here, we use the same basic idea, but make the residual input-dependent: rather than adding a fixed skip connection, we let the fission module attend over the residual states of previous layers and form an input-dependent weighted residual summary. In our implementation, we first construct a lightweight summary $\mathbf{c}'_i = \sum_{l=1}^L \phi^{(l)} \mathbf{c}_i^{(l)}$, where $\phi^{(l)} \in \mathbb{R}$ are learned fixed scalar weights. A selector $S : \mathbb{R}^d \rightarrow \mathbb{R}^L$ then uses this summary to produce input-dependent layer attention weights, which are used to compute the attention residual $\bar{\mathbf{c}}_i$:

$$\omega_i = \text{Softmax}(S(\mathbf{c}'_i)), \quad \bar{\mathbf{c}}_i = \sum_{l=1}^L \omega_i^{(l)} \mathbf{c}_i^{(l)}, \quad \tilde{\mathbf{c}}_i = \mathbf{c}_i^{(L)} + \bar{\mathbf{c}}_i. \quad (2)$$

In this way, $\tilde{\mathbf{c}}_i$, used as input to the modality fission process, allows it to draw selectively on whichever abstraction level is most useful at each step.

Figure 1 illustrates LF²AR (Late Fusion and Late Fission with Attention Residuals), the studied architecture, in the speech setting.

²For potentially useful background on the transformer architecture, please refer to Appendix G.

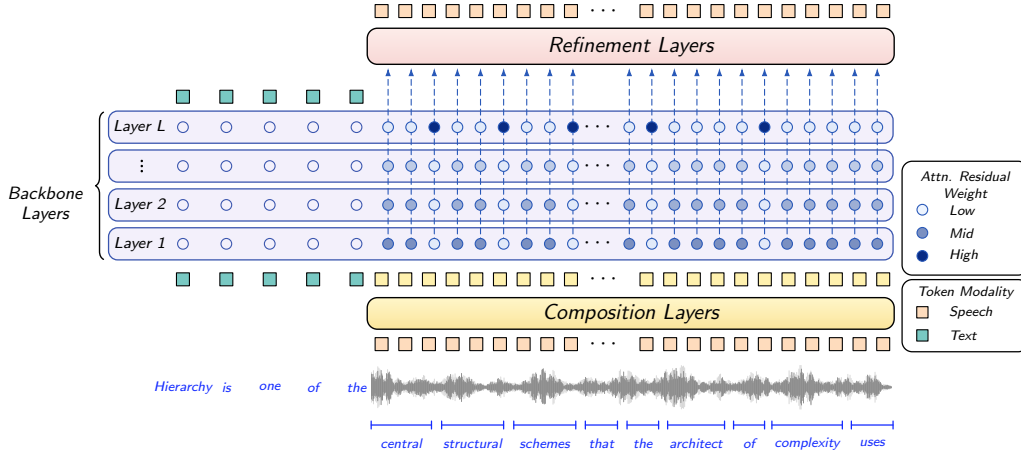


Figure 1: Late Fusion and Late Fission with Attention Residuals (LF²AR), illustrated for a text–speech sequence. Text interfaces directly with the pretrained backbone, while speech representations pass through additional composition and refinement layers. An attention residual allows the refinement module to selectively access residual-stream states across backbone depths. The attention weights illustrate our hypothesized sparse use of depth.

Analysis tools. Our architectural hypotheses are framed in terms of *benefit*: late fusion, late fission, and attention residuals should help multimodal adaptation if they better support cross-modal transfer. To make this concrete, we next introduce a set of layerwise metrics that let us study these benefits in the model’s internal representations.

- **Dataset intrinsic dimensionality.** Following the abstraction-refinement literature, we use the intrinsic dimensionality of the dataset representation manifold as a geometric proxy for feature abstraction and compositionality (Valeriani et al., 2023). For each layer l , let $\mathcal{H}^{(l)} = \{\mathbf{h}_n^{(l)}\}_{n=1}^N$ denote the set of sample representations extracted from that layer. We estimate the intrinsic dimensionality of $\mathcal{H}^{(l)}$, using the *Generalized Ratios Intrinsic Dimension Estimator* (GRIDE) (Denti et al., 2022). Higher intrinsic dimension is interpreted as evidence of a richer, more compositional feature basis.
- **Word-level cross-modal similarity.** Because our multimodal sequences are interleaved at the word level, we can use alignments between text tokens and modality tokens corresponding to the same linguistic unit. Let $\mathcal{B}$ denote the set of aligned word boundaries, and let $\mathbf{h}_{\text{text},j}^{(l)}$ and $\mathbf{h}_{\text{mod},j}^{(l)}$ denote the text and target-modality representations at boundary $j \in \mathcal{B}$ in layer l . We define word-level cross-modal similarity as $\frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \cos(\mathbf{h}_{\text{text},j}^{(l)}, \mathbf{h}_{\text{mod},j}^{(l)})$. This gives a simple measure of cross-modal alignment at the level of linguistic units.
- **BERT alignment.** To measure semanticity, we compare the neighborhood structure of target-modality representations to that induced by a pretrained BERT-style text embedding model. For each layer l , we compute the k -nearest-neighbor graphs of target-modality representations and aligned text semantic embeddings across samples in $\mathcal{H}^{(l)}$, and compute their overlap. Higher values indicate more semantic representations.
- We train TunedLens-style affine probes (Belrose et al., 2025) at every layer. We consider three prediction tasks:
 - **Current-modality-token perplexity** measures how easily the current perceptual unit can be reconstructed from the representation. Lower perplexity indicates that more low-level modality-specific information remains available. Because the model’s unembedding layer is trained for next-token prediction rather than decoding the current modality token, we use a separate learned output projection.

- **Next-modality-token perplexity** measures how far the representation has been refined toward predicting the next fine-grained target-modality unit.
- **Next-text-token perplexity** is measured from target-modality representations at aligned word boundaries. Lower perplexity indicates stronger preservation of the text-like next-token predictions learned during pretraining.

For dataset intrinsic dimensionality, BERT alignment, and next-text-token perplexity, we also report as reference the corresponding values obtained from text inputs.

4 Experimental Setup

Models. We conduct our experiments on SmolLM backbones (Ben allal et al., 2025), available in three sizes—135M, 360M, and 1.7B—which we adapt to process and generate text-as-images or speech. To study our architectural hypotheses, we train model variants ranging from an early fusion/fission model, denoted *Base*, to a late fusion, late fission model with attention residuals, denoted *LF²AR*. To assess both component-wise effects and their interactions, we use complementary ablations: additive ablations from *Base* for pixel-adapted LMs, and subtractive ablations from *LF²AR* for speech-adapted LMs. We implement late fusion and late fission as two decoder layers each, with the same architecture as the backbone, placed before and after it, respectively. In preliminary experiments, additional layers brought little gain in dataset intrinsic dimensionality, so we kept this configuration throughout. We use the MiniLM-L6-v2³ embedding model for the calculation of BERT alignment.

Data. For pixel-adapted LMs, we train on the English subset of FineWiki (Penedo, 2025), comprising 6.6M documents. We render it as images using the renderer of Lotz et al. (2023), yielding 7.04B image patches, together with word-level alignments. For speech-adapted LMs, we train on a collection of public English speech datasets, described in Appendix B, totaling 160k hours of audio. We tokenize them using the speech tokenizer of Hassid et al. (2023), yielding 10.89B speech tokens, and use the forced aligner of Pratap et al. (2024) to obtain word timestamps. All representation-analysis metrics are reported on a held-out set.

Training. All models are trained end-to-end to minimize Equation 1 using the AdamW optimizer (Loshchilov & Hutter, 2019). We train the models adapted from SmolLM 135M and 360M for up to 16B tokens. We use the 1.7B backbone only for speech and train it for up to 32B tokens. We adopt Warmup-Stable-Decay learning-rate schedules, with different peak learning rates for the backbone and the late fusion/fission layers. We tuned the peak learning rates separately for each setup, selecting the largest stable values.

Benchmarks. We evaluate on multiple-choice benchmarks commonly used for LLMs: image-rendered versions of MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), and StoryCloze (Mostafazadeh et al., 2016) for pixel LMs, and the spoken version of StoryCloze (Hassid et al., 2023) for speech LMs. For all benchmarks, we also implement cross-modal variants in which questions and answers may be presented in different modalities. We report the accuracy with which the model assigns higher likelihood to the correct answer.

See Appendix B for more details on the experimental setup.

5 Results

Representation analyses. We analyze the layerwise metrics across the different ablated model variants. For this experiment, we use only the 360M backbone. Figures 2 and 3 present the results for the pixel-adapted and speech-adapted LMs, respectively. In both modalities, *Base* and *LF²AR* score on average the lowest and highest, respectively, in compositionality, as measured by dataset intrinsic dimensionality, word-level cross-modal alignment, and semanticity as measured by BERT alignment. Overall, *Base* also exhibits stronger retention

³<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

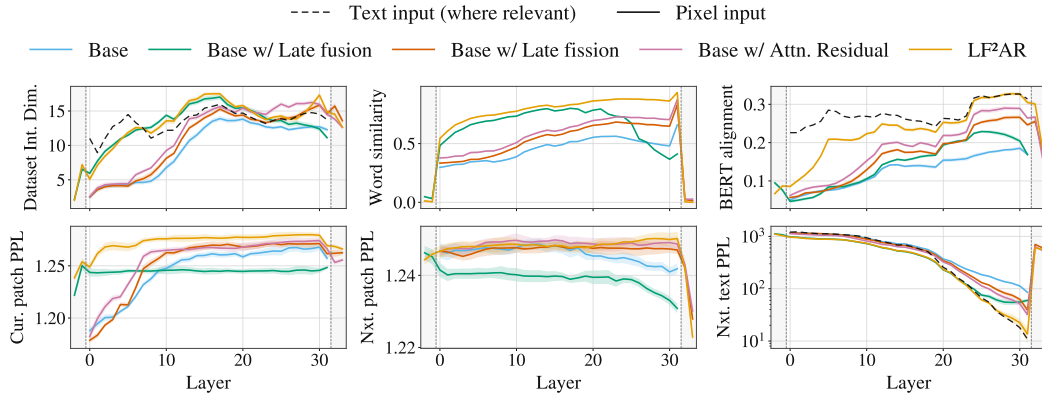


Figure 2: Layerwise analyses for pixel-adapted LMs across additive ablations. Late fusion improves early abstraction and alignment, late fission preserves deeper text-like predictive structure, and attention residuals provide additional gains. Solid curves indicate the means and shaded regions indicate ± 1 standard deviation across five 1,024-sample subsets.

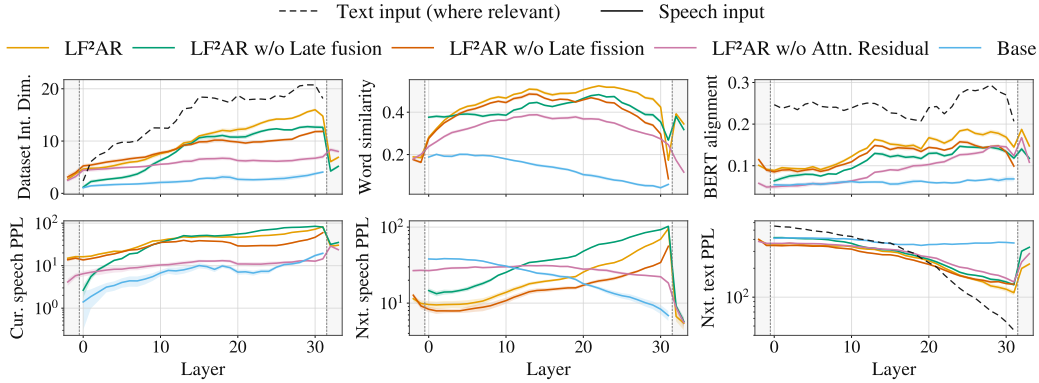


Figure 3: Layerwise analyses for speech-adapted LMs across subtractive ablations. Removing late fusion/fission or attention residuals lowers abstraction, alignment, and semanticity, with attention residuals having the strongest effect on retaining low-level speech detail. Solid curves indicate the means and shaded regions indicate ± 1 standard deviation across five 1,024-sample subsets.

of low-level information and worse language-modeling performance. Notably, *LF²AR* image representations are as semantic and predictive as text representations.

Figure 2 shows that, for pixel adaptation, adding late fusion yields higher compositionality and word-level cross-modal alignment. This effect persists through roughly the first two-thirds of the model depth, after which both metrics drop toward *Base* levels. At that point, next-image-patch perplexity indicates that the model begins to specialize in fine-grained prediction. In the deeper layers, late fusion also yields representations with higher semanticity than *Base* and that are more predictive of language, as shown by next-text-token perplexity. The addition of late fusion also leads to a rapid early loss of low-level detail. Figure 3 shows a similar picture for speech adaptation: removing late fusion produces early-layer representations that are less abstract and less semantic, while also retaining more low-level detail and exhibiting overall worse language-modeling performance.

Key insight: Late fusion induces more text-like early-layer features, leading to more semantic intermediate representations and improved cross-modal transfer.

The clearest effect of late fission is to eliminate the specialization of the deepest layers for fine-grained prediction, which in turn yields higher compositionality, cross-modal alignment, semanticity, and greater use of high-level information predictive of text. In speech, removing

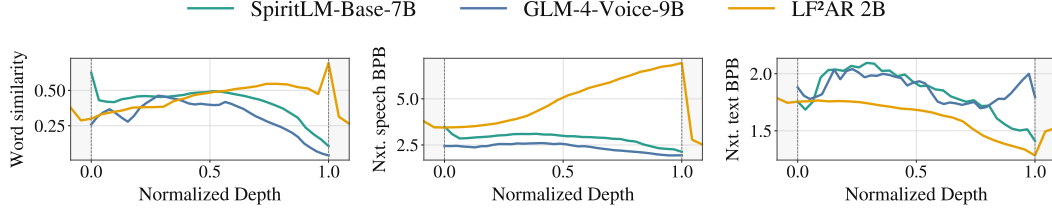


Figure 4: In open-weight speech LLMs with early fission, deeper layers become specialized for fine-grained speech prediction and lose next-word-predictive structure.

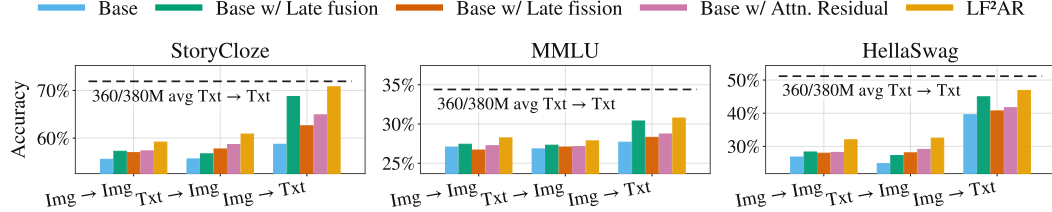


Figure 5: Downstream benchmark accuracy for pixel-adapted LMs across additive ablations from *Base*. Gains are largest when reasoning must be recovered from image input and expressed in text ($\text{Img} \rightarrow \text{Ttxt}$), with late fusion contributing most of the improvement and late fission plus attention residuals adding consistent further gains.

late fission produces the corresponding effect, reducing these metrics in the deepest layers. Interestingly, Figure 4 illustrates the same phenomenon in the open-weight speech-adapted LLMs SpiritLM-7B (Nguyen et al., 2025) and GLM-4-Voice-9B (Zeng et al., 2025), showing that specialization of the deepest layers toward the target modality consistently coincides with the loss of next-word-predictive features. Notably, our adaptation of the 1.7B backbone, *LF²AR* 2B, seems⁴ to achieve stronger next-text-token prediction despite being a substantially smaller model. We note that, in this comparison, we report bits-per-byte (BPB) rather than perplexity to allow a fairer comparison across models with different tokenizers.

Key insight: Late fission prevents the late layers of the pretrained model from being repurposed for fine-grained modeling, preserving pretrained next-word prediction.

Adding attention residuals to late fission in pixel-adapted LMs does not alter the layerwise dynamics of late fission alone, but provides a modest consistent boost in compositionality, cross-modal alignment, and semanticity. In speech, however, removing attention residuals leads to a marked increase in the retention of low-level detail and a corresponding drop in compositionality, cross-modal alignment, and semanticity. Across all subtractive ablations, removing attention residuals has the strongest effect on the retention of low-level detail. We hypothesize that this stronger effect in speech reflects its higher entropy relative to artificially rendered text, which causes greater interference with the backbone’s normal function when propagated through it.

Key insight: Fission with attention residuals removes the need to propagate low-level information throughout the pretrained backbone, thereby allowing higher-level features to form and enabling cross-modal transfer.

For additional experiments on the importance of capacity placement and the dynamic weighting implemented by attention residuals, see Appendices D and E.

Effects on downstream performance. Figures 5 and 11 show that the above representation-level changes translate into consistent downstream gains. For pixel-adapted LMs, improvements are modest when both input and output remain in the image modality, but much

⁴We qualify this statement because, despite reporting bits per byte, different tokenizations induce factorizations of the sequence distribution of varying prediction difficulty.

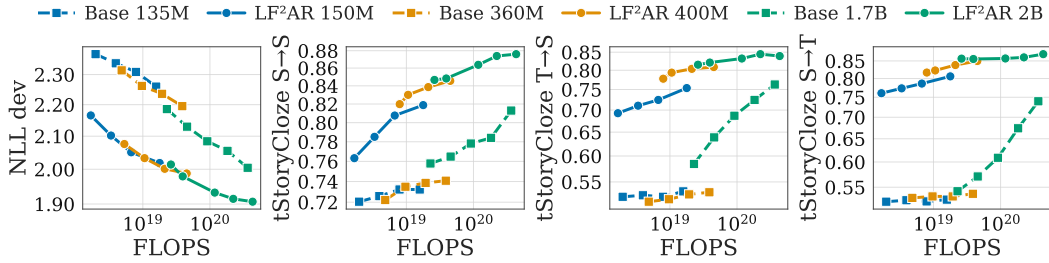


Figure 6: Scaling curves for speech-adapted LMs. LF^2AR improves test negative-log-likelihood (NLL) and spoken StoryCloze performance across compute (in FLOPs) scales.

larger when the model must answer in text. Across StoryCloze, MMLU, and HellaSwag, LF^2AR performs best in the $Img \rightarrow Txt$ setting, with late fusion accounting for most of the gain and late fission plus attention residuals adding further improvements. This matches the layerwise analyses: stronger early abstraction and preserved late text-like refinement are especially useful when textual reasoning must be recovered from perceptual input. The pattern holds for speech LMs. LF^2AR outperforms *Base*, while removing late fusion/fission or attention residuals hurts performance. Late fusion has the largest effect, while late fission and attention residuals provide additional gains, especially in cross-modal settings.

Key insight: Late fusion delivers the largest gains in perceptual understanding, which also improves generation, while late fission is most important for perceptual generation; attention residuals add further gains, especially in cross-modal settings.

Figures 6 and 12 show that these gains persist across compute scale. For pixels, LF^2AR improves over *Base* at both 150M and 375M parameters, with the largest gains on $Img \rightarrow Txt$. For speech adaptation, LF^2AR scales much more favorably. Most notably, the cross-modal $T \rightarrow S$ and $S \rightarrow T$ settings improve sharply with scale for LF^2AR , whereas *Base* remains nearly flat. These architectural gains often exceed what is obtained by scaling the baseline alone: e.g., 150M LF^2AR models are competitive with or stronger than 360M *Base* models, and 400M LF^2AR models consistently outperform 1.7B-scale *Base* in cross-modal transfer.

Key insight: Using late fusion and fission with attention residuals often outperforms parameter scaling alone, and makes cross-modal capabilities appear at smaller scale.

Table 5 shows that speech-adapted LF^2AR models achieve strong performance on StoryCloze, competitive with literature baselines trained with orders of magnitude more compute.

Dynamical depth with attention residuals. Figure 7 shows that attention residuals learn the two-mode pattern hypothesized in Section 3. Most attention mass concentrates on lower layers and the deepest backbone layer, while intermediate layers receive little attention. The attention maps also show that access to the deepest layer is sparse in the sequence dimension. To interpret this sparsity, we measure the correlation between attention spikes and word boundaries by treating attention peaks as boundary predictions and computing an R -value segmentation score (Räsänen et al., 2009). As shown in the right panel of Figure 7, the layers that receive the most attention are also those whose spikes correlate most strongly with word boundaries. This suggests that attention residuals learn to consult high-level textual features sparsely and at a dynamic rate that matches the changes in linguistic units.

Model	t-StoryCloze				Tok/s
	T	S	T→S	S→T	
LF^2AR+	91.0	83.9	79.9	84.6	25.1
LF^2AR+ w/ early exit	91.0	83.3	79.3	85.0	47.7

Table 1: Early-exit decoding for LF^2AR . Skipping deeper layers guided by the attention residual weights preserves performance while increasing throughput $1.9\times$.

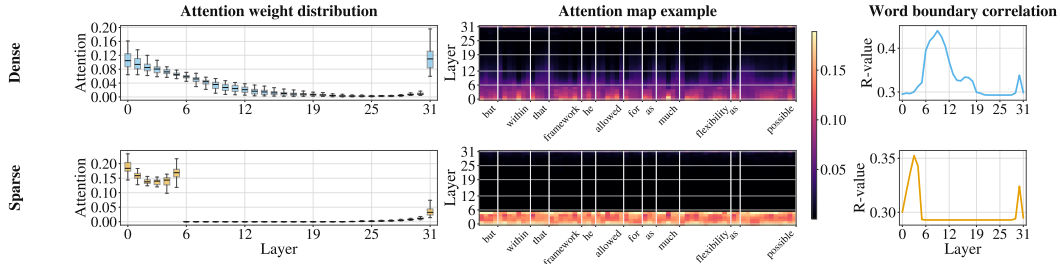


Figure 7: Attention residuals learn sparse, interpretable backbone depth access. **Left:** Layer-attention distributions before/after sparsity tuning. **Middle:** example token-level attention maps. **Right:** word-boundary detection score using attention spikes as predictors.

This phenomenon suggests that much of the deep-backbone computation could be made conditional. To test this, we continue training for roughly 100M tokens, retuning the selector to use a fixed weighted summary of only the first six backbone layers and adding auxiliary losses that encourage sparse depth use, yielding the sparser attention patterns in Figure 7 (bottom). We then train a lightweight router on the same prefix summary to decide whether each token should use a prefix-only approximation or continue through deeper layers. Further details are in Appendix C. Table 1 shows that this preserves performance while increasing throughput by $1.9\times$ under autoregressive decoding without KV caching.

Key insight: Modality fission with attention residuals induces sparse backbone usage, enabling variable depth allocation at inference with minimal performance loss.

6 Limitations

The scope of this paper is restricted in three respects. First, we study perceptual signals—spoken language and rendered text—for which semantic correspondence to text can be controlled closely enough to support word-level alignment and straightforward cross-modal analyses. Whether the same layer dynamics and architectural choices extend to modalities without such near one-to-one correspondence to text, such as natural visual scenes, remains to be established. Second, we study adaptation of a pretrained text LM, whose layerwise organization is established before exposure to another modality. Native multimodal pretraining may instead organize representations and computations across depth differently from the outset. Third, our experiments are limited to models of up to 2B parameters, and extrapolation to substantially larger scales remains uncertain.

Finally, the throughput gains from dynamic depth allocation do not translate directly to deployment in standard autoregressive inference pipelines. Such pipelines assume a uniform computation path and maintain a key-value cache at every layer for each preceding token. Realizing robust serving gains will therefore require specialized inference implementations, as in other conditional-depth methods (Raposo et al., 2024).

7 Conclusions

We argued that multimodal adaptation should account for the layerwise abstraction-refinement dynamics of text language models. From this view, we derived simple architectural hypotheses: finer-grained modalities benefit from extra fine-to-coarse processing before the backbone, extra coarse-to-fine processing after it, and dynamic access to multiple backbone representation levels during generation. Across text-as-images and speech, these design choices consistently improved compositionality, cross-modal alignment, preservation of text-like predictive structure, and downstream accuracy, often with gains that exceeded baseline parameter scaling alone. Attention residuals further induced sparse, interpretable use of deep layers and enabled early-exit decoding with a $1.9\times$ speedup. Taken together, these results show that effective multimodal adaptation benefits from matching the layerwise organization learned during text pretraining.

Acknowledgments

Santiago and Ricard are grateful to the French National Research Agency for its support through grant ANR-20-CE23-0012-01 (MIM), as well as to the Institute of Convergence ILCB, supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). Adel is funded by the Cambridge Trust and is grateful to his former supervisor in Avignon, Professor Yannick Estève, for involving him in this project and providing sponsorship. Other authors also received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 101007666. This work was supported by HPC GENCI-IDRIS under grants AD011014044R1, AD011014044R2, A0161014876, AD011015344, A0161015099, AD011013061R3, and A0161014871.

Finally, we thank the JSALT 2024 workshop and the Center for Language and Speech Processing at Johns Hopkins University for generously hosting the event, where some of the earliest ideas behind this work began to take shape and where fruitful collaborations and friendships emerged in a stimulating and welcoming environment.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. 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=EbMuimAbPbs>.
- Richard Antonello and Emily Cheng. Evidence from fMRI supports a two-phase abstraction process in language models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL <https://openreview.net/forum?id=VZipjF1Bp1>.
- Timo Baumann, Arne Köhn, and Felix Hennig. The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening. *Lang. Resour. Eval.*, 53 (2):303–329, jun 2019. ISSN 1574-020X. doi: 10.1007/s10579-017-9410-y. URL <https://doi.org/10.1007/s10579-017-9410-y>.
- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL <https://arxiv.org/abs/2303.08112>.
- Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Smollm-corpus, 2024. URL <https://huggingface.co/datasets/HuggingFaceTB/smollm-corpus>.
- 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 Fahlgrén, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourier, 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. URL <https://openreview.net/forum?id=3JiC12A14H>.
- Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models, 2024. URL <https://arxiv.org/abs/2405.09818>.
- Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=0fD3iIBh1V>.

- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL <https://arxiv.org/abs/2311.07919>.
- Santiago Cuervo and Ricard Marxer. “scaling properties of speech language models”. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 351–361, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.21. URL <https://aclanthology.org/2024.emnlp-main.21/>.
- Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. Closing the gap between text and speech understanding in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=dDHn03Vhyj>.
- Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. URL <https://arxiv.org/abs/2510.26583>.
- Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, November 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20991-1. URL <https://doi.org/10.1038/s41598-022-20991-1>.
- Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From pixels to words – towards native vision-language primitives at scale. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=DF6udvxuvY>.
- Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=y01KGvd9Bw>.
- Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. The people’s speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL <https://openreview.net/forum?id=R8CwidJ0yT>.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, June 2023.
- Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. DADApY: Distance-based analysis of data-manifolds in python. *Patterns*, 3(10):100589, 2022. doi: 10.1016/j.patter.2022.100589. URL <https://doi.org/10.1016/j.patter.2022.100589>.
- 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=U1HueVjAKr>.
- 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. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

- François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Alexey Karpov, Oliver Jokisch, and Rodmunga Potapova (eds.), *Speech and Computer*, pp. 198–208, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99579-3.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/huh24a.html>.
- Chaoya Jiang, Jia Hongrui, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. MaVEn: An effective multi-granularity hybrid visual encoding framework for multimodal large language model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=fHxmoeKQBh>.
- J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadaiy, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020. doi: 10.1109/ICASSP40776.2020.9052942.
- Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, Yanru Chen, Xin Men, Haiqing Guo, Y. Charles, Haoyu Lu, Lin Sui, Jinguo Zhu, Zaida Zhou, Weiran He, Weixiao Huang, Xinran Xu, Yuzhi Wang, Guokun Lai, Yulun Du, Yuxin Wu, Zhilin Yang, and Xinyu Zhou. Attention residuals, 2026. URL <https://arxiv.org/abs/2603.15031>.
- Kushal Lakhota, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl.a.00430. URL <https://aclanthology.org/2021.tacl-1.79>.
- Frédéric Landragin. Physical, semantic and pragmatic levels for multimodal fusion and fission. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pp. 346–350. Universitätsverlag Tilburg, January 2007. URL <https://shs.hal.science/halshs-00138503>.
- 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 *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. Text rendering strategies for pixel language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10155–10172, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.628. URL <https://aclanthology.org/2023.emnlp-main.628/>.
- Zdravko Marinov, Simon Reiß, David Kersting, Jens Kleesiek, and Rainer Stiefelhagen. Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2283–2293, October 2023.

- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL <https://aclanthology.org/N16-1098/>.
- 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.
- Guilherme Penedo. Finewiki, 2025. URL <https://huggingface.co/datasets/HuggingFaceFW/finewiki>. Source: Wikimedia Enterprise Snapshot API (<https://api.enterprise.wikimedia.com/v2/snapshots>). Text licensed under CC BY-SA 4.0 with attribution to Wikipedia contributors.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024. URL <http://jmlr.org/papers/v25/23-1318.html>.
- Enrique Queipo-de-Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael M. Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=c5TFhCJ6fs>.
- David Raposo, Samuel Ritter, Blake A. Richards, Timothy P. Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024. doi: 10.48550/arXiv.2404.02258.
- Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altsaar. An improved speech segmentation quality measure: The r-value. In *Tenth Annual Conference of the International Speech Communication Association (Interspeech 2009)*, pp. 1851–1854, 2009. URL http://www.isca-speech.org/archive/interspeech.2009/i09_1851.html. Funded by EU FP6 FET project ACORNS (FP6-034362).
- Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=FkSp8VW8RjH>.
- Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and LILI YU. LMFusion: Adapting pretrained language models for multimodal generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=Kc1WTxZbrP>.
- Mustafa Shukor, Enrico Fini, Victor Guilherme Turrissi 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 (ICCV)*, pp. 12–23, October 2025.

- 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>.
- Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. SEAM: Semantically equivalent across modalities benchmark for vision-language models. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=1I4LgGv4sX>.
- Liang-Hsuan Tseng, Yi-Chang Chen, Kuan Yi Lee, Da shan Shiu, and Hung yi Lee. TASTE: Text-aligned speech tokenization and embedding for spoken language modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=6STb8DauN1>.
- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=cCYvakU5Ek>.
- Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. Too late to recall: Explaining the two-hop problem in multimodal knowledge retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=qeL8fi8GS7>.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL <https://aclanthology.org/2021.acl-long.80>.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 12966–12977, June 2025.
- Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL <https://arxiv.org/abs/2408.16725>.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL <https://arxiv.org/abs/2503.20215>.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2024. URL <https://api.semanticscholar.org/CorpusID:274514545>.
- Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Connecting speech encoder and large language model for asr. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12637–12641, 2024. doi: 10.1109/ICASSP48485.2024.10445874.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, shengmin jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=3tukjsVyrE>.

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*, 2024. URL <https://openreview.net/forum?id=nqWaya7hiX>.

Xingjian Zhao, Zhe Xu, LuoZhijie Jin, Yang Wang, Hanfu Chen, Yaozhou Jiang, Ke Chen, Ruixiao Li, Mingshu Chen, Ruiming Wang, Wenbo Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. Towards true speech-to-speech models without text guidance. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=zjaV5zmlkl>.

A Related Work

Many prior works adopt late-fusion architectures to improve modality adaptation. In the speech domain, relevant examples include Yu et al. (2024); Chu et al. (2023); Tang et al. (2024). More recently, several models have focused on decoupling speech generation from the text LM backbone through additional speech-specific output modules (Xu et al., 2025) or modality-specific forward passes (Xie & Wu, 2024), an idea aligned with the intuition behind our late-fission approach: non-textual modalities may also benefit from additional output-side processing. Unlike these works, we adopt a much simpler approach to modality fusion and fission and motivate it through the abstraction-refinement phenomenon observed in transformer models, as described in Section 3. We also do not rely on paired text conditioning for speech generation.

Perhaps most relevant is the work of Zhao et al. (2026), which studies a modality-based layer-splitting architecture to improve the coupling of speech with pretrained text LLMs and similarly reports representation-level and downstream benefits, albeit using a less comprehensive set of measures. Our attention residuals can be viewed as enabling an input-dependent, dynamic form of layer selection rather than relying on a fixed layer-splitting strategy. We further show that this dynamic selection enables variable-depth allocation and can yield decoding speedups. More broadly, relative to prior work, our representation analyses provide additional insight into the roles and effects of different fusion and fission strategies in multimodal adaptation.

B Additional Setup Details

Training. All models use a weight decay of 0.1. We use a peak learning rate of $3e-4$ for backbone parameters in the 135M and 360M models, and $1e-4$ for the 1.7B model. The learning rate for non-backbone parameters is uniformly set to $10\times$ that of the backbone parameters. We use a linear warmup of 100 steps and linearly decay the learning rate over the final 20% of training. We use a batch size of 1M tokens with a context length of 2048 tokens. Each batch contains equal proportions of target-modality-only, text-only, and interleaved text-and-target-modality sequences. To produce interleaved cross-modal sequences, we segment each utterance into subsequences and interleave spans of random length at runtime: 10–30 words for text and pixel segments and 5–15 words for speech segments. For the affine probes, we use a learning rate of 1.0 and linearly decay it to zero over 1000 training steps, with a batch size of 256 samples.

Data. For speech-adapted LMs, we use LibriSpeech (Panayotov et al., 2015), LibriLight (Kahn et al., 2020), SWC (Baumann et al., 2019), Tedlium (Hernandez et al., 2018), People (Galvez et al., 2021), Vox Populi (Wang et al., 2021), and sTinyStories (Cuervo & Marxer, 2024). As noted above, we also train the models on text-only data. For this, we use a 12-billion-token subset of the SmolLM corpus (Ben Allal et al., 2024), with a data distribution matching that used to pretrain the SmolLM models⁵.

⁵As reported in <https://github.com/huggingface/smollm/blob/main/pre-training/>

Representation analysis. We compute all metrics over five subsets of 1,024 samples, independently sampled with replacement, with each sample consisting of 20 words in the corresponding modality. For pixel models, we use a held-out set from FineWiki, amounting to 0.1% of the training set. For speech, we use data from the StoryCloze benchmark. Curves report the mean across the five subsets, with shaded regions indicating ± 1 standard deviation. For probe-based perplexity metrics, we additionally train each probe with three random seeds and compute the mean and standard deviation jointly across sample subsets and probe seeds. To compute intrinsic dimensionality and BERT alignment, we pool one representation per sample: for intrinsic dimensionality, we take the final representation in the sequence; for BERT alignment, we use average pooling. We use average pooling for BERT alignment because we found that it yields more discriminative results. Since transformer representations exhibit large activation outliers, when computing intrinsic dimensionality and BERT alignment we truncate feature elements (i.e., individual activations) that exceed the 99th percentile across the dataset, following Huh et al. (2024). To estimate intrinsic dimensionality, we use the GRIDE estimator (Denti et al., 2022) implemented in dadapy (Glielmo et al., 2022), following the procedure described by Cheng et al. (2025).

Benchmarking. We synthesize target-modality versions of the benchmark text data using the Kokoro-TTS model⁶ with the af-heart voice, and the same text renderer used to generate the text-as-images training data. All benchmarks are multiple choice. For each task, the model estimates the normalized log probability of each answer given the context, and accuracy is computed by checking whether the highest-scoring option matches the gold answer. Table B presents the prompt templates used for each task.

StoryCloze

<story prefix> <answer>

MMLU

The following are multiple choice questions (with answers) about <topic>.

Question: <question>

Answer: <answer>

HellaSwag

<context> <answer>

Table 2: Evaluation task prompts. Placeholders are shown in monospace. Blue highlights denote text/target-modality inputs, and orange highlights denote continuations where likelihood is evaluated.

⁶<https://github.com/hexgrad/kokoro-82M>

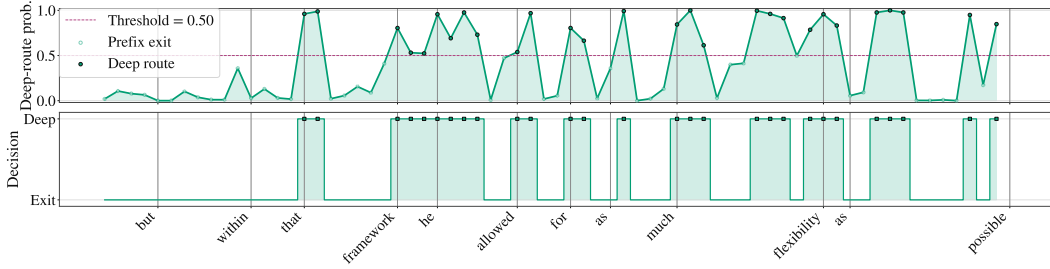


Figure 8: Example router decisions for the sparsity-tuned 360M speech LF^2AR model. The router, conditioned on a six-layer prefix summary, assigns each token a probability of continuing through the deeper backbone layers (top) and yields a corresponding hard early-exit versus deep-route decision (bottom). Deep routing is used only sparsely along the sequence, illustrating how the model allocates additional compute selectively.

C Sparse Selector Tuning and Early Exit

For the early-exit experiment in Section 5, we start from the 360M speech LF^2AR model and continue training for roughly 100M speech tokens using speech-only batches. During this stage, we freeze the pretrained backbone and update only the speech-specific input/output layers, the fixed layer weights $\phi^{(l)}$, the selector S , and, in the routing stage, a small linear router head.

To make sparse depth decisions available before the full backbone has run, we replace the selector controller input in Equation 2 by a fixed weighted summary of only the first six backbone layers. Concretely, if $\mathbf{c}_i^{(1)}, \dots, \mathbf{c}_i^{(6)}$ are the first six residual-stream states, we define

$$\mathbf{c}'_i = \sum_{l=1}^6 \tilde{\phi}^{(l)} \mathbf{c}_i^{(l)},$$

where $\tilde{\phi}^{(l)}$ are the learned fixed layer weights restricted and renormalized to those six layers.

We then add a sparse-selector penalty to the language-modeling objective ($\mathcal{L}_{\text{mm}}$):

$$\mathcal{L} = \mathcal{L}_{\text{mm}} + \lambda_D \sum_{l=1}^L \frac{l-1}{L-1} \omega_i^{(l)} + \lambda_{\text{KL}} \text{KL}(\omega_i \| \mathbf{q}),$$

averaged over speech positions. We use $\lambda_D = 0.05$, and $\lambda_{\text{KL}} = 0.02$, with a linear warmup over the first 20 optimization steps. The target distribution $\mathbf{q}$ assigns 0.8 total mass to the first six layers in proportion to the renormalized fixed layer weights, 0.2 mass to the final backbone layer, and negligible mass elsewhere.

To obtain actual conditional computation at inference time, we train a lightweight router $r(\mathbf{c}'_i)$ with a binary cross-entropy loss of weight 1.0. The router target is 1 iff, in a frozen dense teacher model, the selector peaks on the final backbone layer at position i . At inference time, the first six backbone layers are always evaluated. If $\sigma(r(\mathbf{c}'_i)) < 0.5$, the model uses a prefix-only proxy for the audio context; otherwise it continues through the deeper backbone layers. Figure 8 illustrates the router predictions for a given sample.

D Matched-Capacity Layer-Placement Controls

A potential alternative explanation for the gains of LF^2AR is that they arise from its additional modality-specific capacity rather than from where that capacity is placed. We therefore conduct a matched-capacity placement ablation in the 400M speech setup. All variants use the same pretrained 360M backbone, the same four newly initialized modality-specific decoder layers, the same attention-residual mechanism, and the same training budget. They differ only in the placement of the additional layers: all four layers before the

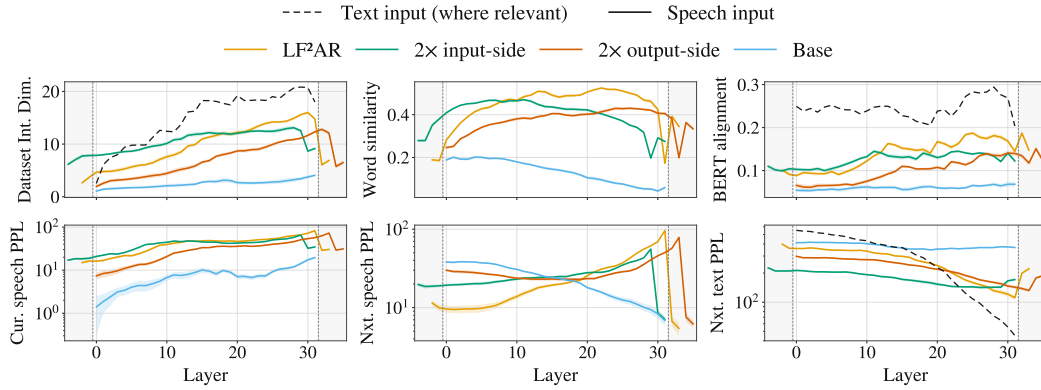


Figure 9: Layerwise analyses for matched-capacity placement controls. Input-side placement primarily improves early abstraction and word-level alignment, whereas output-side placement better preserves modality- and text-predictive structure in later layers. Splitting the additional depth across both interfaces, as in LF²AR, yields the strongest overall abstraction, cross-modal alignment, semantic alignment, and reuse of text-learned predictive structure. Solid curves indicate the means and shaded regions indicate ± 1 standard deviation across five 1,024-sample subsets.

Model	T $\uparrow$	S $\uparrow$	T $\rightarrow$ S $\uparrow$	S $\rightarrow$ T $\uparrow$
2 $\times$ input-side	90.9	82.4	76.8	81.2
2 $\times$ output-side	90.6	80.7	70.6	82.8
LF ² AR	91.3	84.6	80.9	85.0
<i>p</i> -value vs. best placement control	0.54	0.010	1.4×10^{-5}	0.0096

Table 3: Downstream performance of the matched-capacity placement controls on tStoryCloze. All variants have the same newly initialized modality-specific capacity. We report two-sided two-proportion tests comparing LF²AR with the strongest placement control in each evaluation setting.

backbone (2 $\times$ input-side), all four layers after the backbone (2 $\times$ output-side), or two layers at each interface as in LF²AR.

Figure 9 shows that the two controls produce the directional effects expected from their placement. Concentrating the additional computation before the backbone produces stronger early abstraction and word-level alignment than concentrating it after the backbone. Conversely, the output-side control provides better next-speech and next-text prediction, consistent with allocating more modality-specific computation to refinement. Neither placement alone reproduces the complete LF²AR profile. Splitting the layers across the two interfaces combines stronger fine-to-coarse processing before the backbone with improved coarse-to-fine refinement after it.

As shown in Table 3, LF²AR performs best in all four modality configurations. The difference is small when both input and output are text, where the pretrained backbone already operates in its native modality, but is larger in the speech and cross-modal settings. In particular, the improvements in S, T $\rightarrow$ S, and S $\rightarrow$ T are significant relative to the strongest matched-capacity placement control. These results indicate that the gains are not explained solely by the number of additional modality-specific layers: where the added computation is placed materially affects cross-modal transfer.

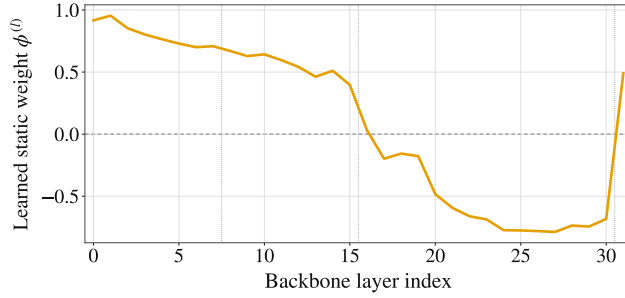


Figure 10: Learned fixed summary weights in the main 400M LF²AR model. The weights $\phi^{(l)}$ place large positive weight on early backbone states, suppress intermediate and deep layers, and return to a positive value at the final layer. These are the fixed weights used to construct the input to the attention-residual selector, rather than the input-dependent selector weights ω_i themselves.

Model	T $\uparrow$	S $\uparrow$	T $\rightarrow$ S $\uparrow$	S $\rightarrow$ T $\uparrow$
LF ² AR-400M	91.3	84.6	80.9	85.0
Static residual-400M	90.8	84.0	80.1	83.9

Table 4: Dynamic versus static residual access on tStoryCloze. The static-residual control has access to the same intermediate backbone states but cannot select among them as a function of the input. Dynamic attention residuals perform best in every modality configuration.

E Fixed Summary Weights and a Static-Residual Alternative

The attention-residual mechanism first constructs the lightweight summary

$$\mathbf{c}'_i = \sum_{l=1}^L \phi^{(l)} \mathbf{c}_i^{(l)}, \quad (3)$$

where the learned scalars $\phi^{(l)}$ are fixed across inputs. A selector then uses $\mathbf{c}'_i$ to produce the input-dependent weights ω_i used by the fission module. Figure 10 shows the learned fixed summary weights from the 400M LF²AR model.

The learned profile exhibits the two-mode organization expected from the abstraction-refinement view. Early states provide the selector with low-level and local perceptual information, while the positive final-layer weight provides access to the backbone’s most text-predictive representation. Intermediate and deep layers are comparatively suppressed.

To determine whether fixed access to multiple representation levels is sufficient, we additionally train a static-residual control. This model uses the same pretrained backbone, modality-specific capacity, and training budget as LF²AR, but removes the input-dependent selector and supplies the summary in Equation 3 as the residual to the fission module. The comparison therefore isolates the contribution of dynamically selecting representation depth from the more general benefit of accessing intermediate backbone states.

Table 4 shows that fixed residual access recovers part of the benefit but remains consistently weaker than input-dependent attention residuals. The largest difference occurs in S $\rightarrow$ T, where successful transfer requires recovering text-like predictive structure from speech representations. Moreover, unlike dynamic attention residuals, the static alternative cannot support input-dependent depth selection and therefore does not provide the variable-compute early-exit mechanism.

F Additional Figures and Tables

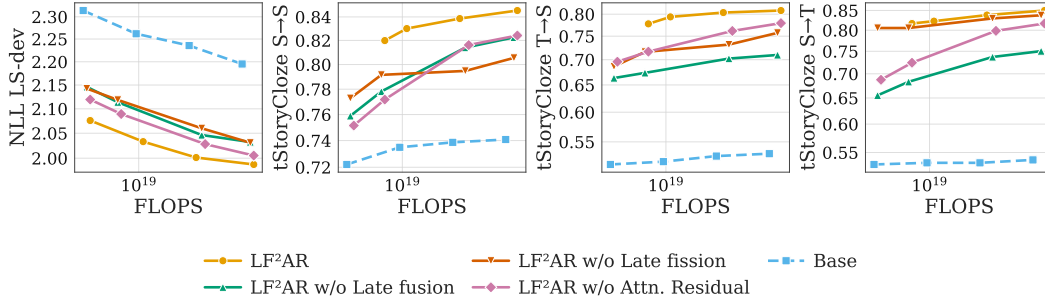


Figure 11: Downstream benchmark performance for speech-adapted LMs across subtractive ablations from LF^2AR . Late fusion has the largest individual effect, while late fission and attention residuals provide additional gains, especially in cross-modal settings.

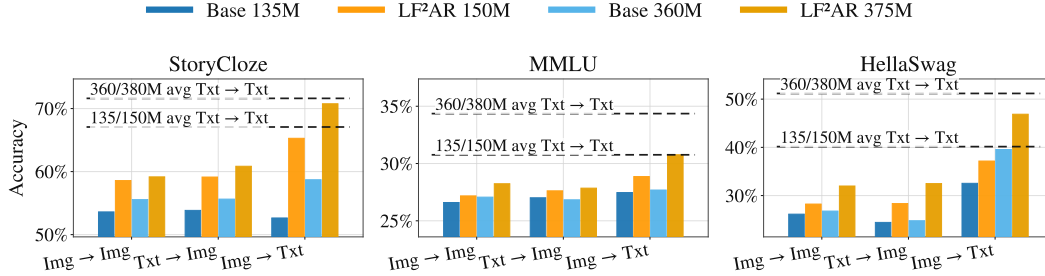


Figure 12: Scaling curves for pixel-adapted LMs. LF^2AR outperforms Base at both scales, with the largest gains in $Img \rightarrow Txt$, indicating that the architectural changes improve transfer efficiency beyond parameter scaling alone.

Model	Params.	Tokens	tStoryCloze				sStoryCloze				MMLU	
			T	S	T→S	S→T	T	S	T→S	S→T	T (adapted/base)	
SpiritLM-Base-7B	7B	~175B	95.8	90.5	78.6	94.3	74.0	66.3	64.7	71.7	37.7 / 39.0	
GLM-4-Voice 1.5B	1.5B	1T	—	77.5	81.4	90.1	—	55.4	58.6	64.0	—	
GLM-4-Voice 9B	9B	1T	—	83.0	85.0	93.6	—	62.4	63.2	76.3	—	
LF^2AR -150M	150M	16B	88.4	82.0	75.2	81.0	64.1	55.0	58.8	58.4	30.0 / 30.2	
LF^2AR -400M	400M	16B	91.3	84.6	80.9	85.0	68.4	57.5	62.3	62.1	34.0 / 34.2	
LF^2AR -2B	2B	32B	92.6	87.6	84.3	87.1	73.6	61.4	64.2	64.2	40.1 / 40.0	

Table 5: Comparison against literature baselines on downstream evaluations. **Best results** are bold and underlined; **second best** are bold. For MMLU, “best” denotes the model with the smallest post-adaptation degradation relative to its text-only base model.

G Transformer Background

We briefly fix terminology for the transformer architecture used throughout the paper. Our models are causal decoder-only transformers that process a sequence of discrete tokens from a vocabulary $\mathcal{V}$ autoregressively through a stack of layers. Let $c_i^{(l)} \in \mathbb{R}^d$ denote the representation at sequence position i after layer l . We refer to the sequence $\{c_i^{(l)}\}_{l=0}^L$ as the *residual stream* at position i . Each decoder layer updates the current residual-stream state by adding the output of its submodules to the running representation inherited from previous layers. Ignoring normalization details, a layer can be written as

$$\hat{c}_i^{(l)} = c_i^{(l-1)} + \text{Attn}^{(l)}(c_{\leq i}^{(l-1)}), \quad (4)$$

$$c_i^{(l)} = \hat{c}_i^{(l)} + \text{MLP}^{(l)}(\hat{c}_i^{(l)}). \quad (5)$$

Thus, rather than replacing the previous representation, each layer incrementally refines a shared state. This residual-stream view is useful because intermediate layers can be understood as exposing different levels of abstraction of the same token position. In an autoregressive language model, the top-layer representation $c_i^{(L)}$ is projected to logits over the next token, defining a distribution over $\mathcal{V}$ conditioned on the preceding context. In this paper, we use the terminology *backbone* to refer to this stack of pretrained transformer layers.

Author contributions

Santiago proposed and led the project, defining the hypotheses, model, and experimental design, developed most of the training, modeling, evaluation, and analysis codebase, and led the writing of the manuscript. Adel contributed to speech data preprocessing, synthesis, and evaluation, while also contributing model-design insights. Ricard provided supervision throughout and contributed to the data pipeline, experimental codebase, and analysis. Yanis assisted with experiment implementation, execution, and discussions. Phil contributed to the manuscript and discussions, as well as to Adel’s supervision.. The other authors contributed to early-stage discussions and supervision during the JSALT 2024 workshop.