WHY DOES NON-TEXT TOKEN NORM INFLATION EMERGE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal LLMs improve at understanding images and audio, yet a modality gap between non-text and text remains. We define *non-text token norm inflation* as an increase in non-text token L2 norms during modality alignment, with these norms exceeding text token norms in early decoder layers. To examine its emergence, we track non-text norms during alignment. We then test its information-preserving role by replacing non-text tokens with inputs of varying norms but fixed directions. We assess the effects of mitigating norm inflation using *Text-Anchored Norm Regularization (TAN)* during Qwen3.5-4B audio alignment. Our findings are threefold. First, early norm inflation is driven by increases in common-component norm. Second, we find an early decoder-layer range where replacing subsequent non-text tokens with inputs preserves at least 90% of originally correct predictions. Finally, TAN applied through early decoder layers yields lower accuracy after replacement than without TAN. TAN achieves higher inter-modal cosine similarity and improves performance across tasks, with a 9.9%p gain on MMAU. These findings support the information-preserving role of norm inflation. The results suggest that mitigating norm inflation may promote earlier formation of cross-modal representations and thereby improve task performance.

1 INTRODUCTION

Multimodal large language models (MLLMs) extend LLMs to images and audio and continue to improve at understanding these inputs (Lu et al., 2025; Gemma Team, 2026; Qwen Team, 2026; Ghosh et al., 2026). However, a modality gap between non-text and text tokens persists after modality alignment (Shukor & Cord, 2024; Xiang et al., 2025). The modality gap in contrastive learning can arise from encoder initialization and persist under an objective that pulls matched representations together while pushing unmatched ones apart (Liang et al., 2022). By contrast, representative MLLMs learn modality alignment through next-token prediction (NTP), without an explicit repulsive term (Liu et al., 2023; Tang et al., 2024). We therefore examine how differences between non-text and text representations develop and persist during NTP training.

Prior work on the modality gap in MLLMs has reported differences in hidden-state direction, distance, and norm (Shukor & Cord, 2024; Xiang et al., 2025; Li et al., 2026). In particular, early layers of pre-norm vision MLLMs exhibit a *norm discrepancy*: non-text tokens have larger L2 norms than text tokens (Li et al., 2026). Although pre-norm architectures normalize sublayer inputs, the residual path retains each token's original magnitude. Token norms can therefore affect how much subsequent updates change representation direction. We investigate the emergence of norm discrepancy during alignment training, the role of large norms in preserving decoder inputs, and the effects of mitigating norm inflation on modality alignment.

We first reduce initial non-text norms and track how they change during vision and audio modality alignment. In both modalities, non-text norms increase again and exceed text norms in the early decoder. We call this increase during training *non-text token norm inflation*. To examine this increase, we decompose non-text representations into a common component, the mean across inputs, and an individual component, the deviation from that mean. The increase in common-component norm accounts for early inflation in modality alignment. The early increase in common-component norm during projector training suggests that shared adjustments make encoder representations compatible with the decoder and contribute to norm inflation.

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.

We hypothesize that norm inflation helps preserve the input representations produced by the encoder and projector. This is because large non-text norms can limit the changes in direction caused by decoder updates (Li et al., 2026). We test this hypothesis through *input replacement experiments*, measuring accuracy after replacing later non-text tokens with inputs at the same positions. In vision and audio models, unscaled inputs retain at least 90% of originally correct answers over a range of early decoder depths. We further vary input norms at fixed directions and relate accuracy differences to changes in decoder representations across models. These comparisons support the role of norm inflation in preserving the input representations themselves.

We next examine how mitigating norm inflation affects input preservation and representation learning during audio alignment. We connect an audio encoder to Qwen3.5-4B, pretrained for images and text, and apply Text-Anchored Norm Regularization (TAN). TAN aims to match the mean and standard deviation of non-text norms to text references. Applied through early decoder layers, TAN lowers accuracy after replacement with unscaled inputs relative to standard NTP training. TAN also yields higher inter-modal cosine similarity than standard NTP training, with high values appearing at shallower decoder depths. We interpret the reduced input preservation and higher similarity under TAN as earlier joint audio–text representation formation than under standard NTP training. These changes accompany performance gains across tasks including transcription and audio understanding; for example, MMAU accuracy rises by 9.9%p. The task gains suggest that the joint representations help the pretrained decoder apply its capabilities to audio inputs.

The main contributions of this paper are threefold. First, we characterize the emergence of norm inflation, which occurs even from reduced initial norms, with increases in common-component norm contributing substantially to early inflation (Section 4.1). Second, using token replacement with norm and direction controls, we provide evidence for the role of norm inflation in preserving input representations within the decoder (Section 4.2). Finally, in Qwen3.5-4B audio alignment, we assess the effects of mitigating norm inflation with TAN through early decoder layers: input preservation decreases, inter-modal similarity increases, and task performance improves (Section 4.3). We thus explain norm inflation during modality alignment through its role in preserving decoder-compatible inputs. The audio results suggest that reducing input preservation can facilitate the transformation of non-text representations with text and help joint representations form earlier.

2 Related Work

Modality alignment in MLLMs connects pretrained non-text encoders to an LLM that generates text conditioned on their outputs. Models such as Flamingo, LLaVA, and OneLLM use cross-attention or projection modules for this purpose (Alayrac et al., 2022; Liu et al., 2023; 2024; Han et al., 2024). Early modality-gap studies focused on contrastively trained image–text representations. CLIP is a prominent contrastively trained image–text model (Radford et al., 2021). Liang et al. (2022) showed that encoder initialization can create a gap that contrastive optimization sustains. Although many generative MLLMs learn text generation from non-text inputs through next-token prediction without explicitly pushing unmatched pairs apart, modality differences persist within their decoders (Liu et al., 2023; Tang et al., 2024). Shukor & Cord (2024) found higher cosine similarity among same-modality mean representations than between non-text and text means. They also reported larger non-text than text norms, smaller directional changes for non-text tokens across decoder layers, and increasing non-text norms during mapping-module training. Xiang et al. (2025) showed that cosine similarity and Euclidean distance between non-text and text mean representations can increase together with decoder depth. In pre-norm MLLMs, Li et al. (2026) interpreted limited directional changes at large non-text norms as delaying inter-modal integration and proposed post-projector norm adjustment. We focus on the emergence of norm inflation during modality alignment from the perspective of preserving decoder-compatible inputs. Appendix A details the broader alignment methods and representation analyses.

3 Methods

We consider pre-norm LLM decoders processing non-text and text tokens jointly. A token is a vector at a decoder sequence position. Non-text input tokens are encoder outputs projected into the LLM decoder's input space. Each decoder sublayer normalizes the input sequence H tokenwise,

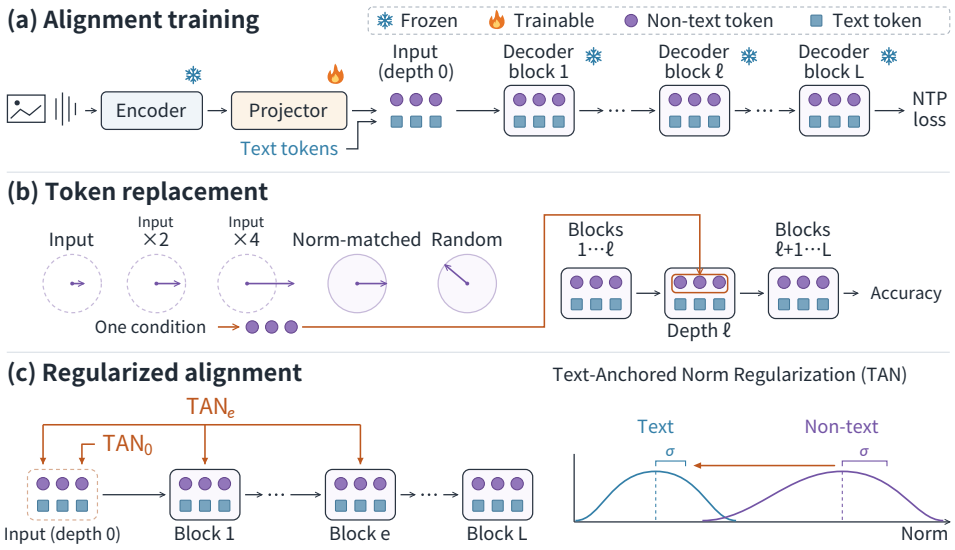


Figure 1: Methods overview. (a) Token and component norms are tracked during NTP projector training. (b) Non-text tokens at depth ℓ are replaced using input scaling, Norm-matched, or Random, retaining text states and subsequent computation. (c) TAN matches non-text norm means and standard deviations to text references: TAN_0 at depth 0 and TAN_e at depths 0–e.

computes an update, and adds it to H: $H^+ = H + F(\mathrm{Norm}(H))$. Here, F denotes attention or an FFN. Depth $\ell = 0$ denotes the input, and $\ell \geq 1$ the output of block ℓ. All norms are L2 norms. Section 3.1 tracks the emergence of norm inflation during alignment, Section 3.2 tests its information-preserving role through token replacement, and Section 3.3 examines the effects of mitigating norm inflation through norm regularization.

3.1 Tracking the Emergence of Norm Inflation

Alignment training. We track norm inflation during projector training in Qwen3.5-4B vision and Qwen2-Audio-7B, keeping the pretrained encoder and decoder frozen. Figure 1(a) shows the training and measurement setup. Each model's projector is reinitialized before training. **Baseline** uses NTP loss with default initialization; **Reduced initial norm** starts NTP training at text-level norms. Appendix B provides data and optimization settings.

Measurements. At each checkpoint and depth, we average token norms within examples and then equally across a fixed set of probes. For non-text token $h_{x,i}$ at position i in example x, we separate the common component $c = \mathbb{E}_x \mathbb{E}_{i|x} h_{x,i}$ from the individual component $h_{x,i} - c$. The mean squared norm decomposes into $\|c\|_2^2 + \mathbb{E}\|h - c\|_2^2$. Over a training interval, each component's contribution is the increase in its term divided by the total increase. Component trajectories show the square roots of these terms: common-component L2 norm and individual-component RMS norm. Appendix D gives the derivation and calculations. At vision Baseline checkpoints, we also test whether non-text input tokens from different images are interchangeable. We replace their directions with those of non-text input tokens from another image or with random directions, retaining the original token norms, and compare NTP loss with that of the original input. Appendix D.1 details this cross-input comparison.

3.2 Testing the Role of Norm Inflation

Token replacement. We test whether inputs can replace later non-text tokens while sustaining predictions. The pretrained models are Qwen3.5-4B and 27B vision, Phi-4-multimodal vision, and Qwen2-Audio-7B. **Original** retains the model's computed tokens. Each intervention replaces tokens at all non-text positions at one depth, retaining text states and subsequent computation, as illustrated in Figure 1(b). **Input** replaces each non-text token $h_{x,i}^{(\ell)}$ with the input $h_{x,i}^{(0)}$ at the same position.

Input $\times 2$ and **Input** $\times 4$ multiply that input by 2 and 4. **Norm-matched** combines the input direction with the norm of the token being replaced:

$$\widetilde{h}_{x,i}^{(\ell)} = \|h_{x,i}^{(\ell)}\|_2 \frac{h_{x,i}^{(0)}}{\|h_{x,i}^{(0)}\|_2}. \tag{1}$$

Random independently samples a standard Gaussian vector for each token, normalizes it, and scales it to the norm of the token being replaced. Comparing Norm-matched with Random isolates the contribution of input direction at the same norm.

Evaluation. We measure accuracy after replacement on 300 items that the model originally answers correctly, with answers determined by the highest-logit option symbol. Each model's cohort is fixed across replacement conditions and depths. Appendix C describes selection and validation. Under Original, **input cosine similarity** compares each non-text token with its depth-0 representation, averaging within examples. **Inter-modal cosine similarity** compares each example's mean non-text and text vectors at the same depth; text comprises the prompt and question. Both measures weight examples equally.

3.3 Assessing the Effects of Mitigating Norm Inflation

Regularized alignment. We examine the effects of mitigating norm inflation by training an audio projector with **Text-Anchored Norm Regularization (TAN)**. The newly initialized linear projector connects the Qwen2.5-Omni-3B audio encoder to the Qwen3.5-4B decoder. Both pretrained components remain frozen during projector training. TAN regularizes NTP training to match the mean and standard deviation of non-text norms to text references at the same depth, as illustrated in Figure 1(c). Consistent with this objective, we scale non-text inputs at initialization, preserving their directions and matching their median norm to the median text norm. For non-text statistics $(\mu_{\mathrm{NT}}^{\ell}, \sigma_{\mathrm{NT}}^{\ell})$ and text references $(\mu_{\mathrm{T}}^{\ell}, \sigma_{\mathrm{T}}^{\ell})$,

$$\mathcal{R}_\ell = \frac{1}{2}\left[\left(\frac{\mu_{\mathrm{NT}}^{\ell} - \mu_{\mathrm{T}}^{\ell}}{\mu_{\mathrm{T}}^{\ell}}\right)^2 + \left(\frac{\sigma_{\mathrm{NT}}^{\ell} - \sigma_{\mathrm{T}}^{\ell}}{\mu_{\mathrm{T}}^{\ell}}\right)^2\right]. \tag{2}$$

TAN$_0$ regularizes the input, and **TAN**$_e$ regularizes the input through depth e:

$$\mathcal{L}_{\mathrm{TAN}_0} = \mathcal{L}_{\mathrm{NTP}} + \lambda \mathcal{R}_0, \qquad \mathcal{L}_{\mathrm{TAN}_e} = \mathcal{L}_{\mathrm{NTP}} + \frac{\lambda}{e+1}\sum_{\ell=0}^{e} \mathcal{R}_\ell. \tag{3}$$

We use $\lambda = 0.1$ for both TAN variants and $e = 11$ for TAN$_e$. Appendix B specifies text references, initialization, training settings, and the choice of e.

Comparisons and evaluation. We compare four training conditions: Baseline, Initialization only, TAN$_0$, and TAN$_e$. Baseline uses NTP loss with default initialization. An **Initialization only** ablation shares TAN's initialization and learning rate but omits the norm term. After projector training, we freeze the projector and train decoder LoRA with NTP (Hu et al., 2022) in all four conditions. We track norms during projector training; replacement and task evaluations use the final LoRA models. We compare accuracy before and after replacement on 300 items that all four models originally answer correctly. On the same items, we also compute the cosine similarities defined in Section 3.2. Task performance is evaluated from generated responses using word error rate (WER) on LibriSpeech test-clean and LibriMix, QA F1 on SpokenSQuAD, and accuracy on MMAU and MMAU-Pro (Panayotov et al., 2015; Cosentino et al., 2020; Lee et al., 2018; Sakshi et al., 2025; Kumar et al., 2026). Appendix F gives all tasks, scoring rules, and denominators.

4 Results and Discussion

We first examine the emergence of norm inflation by tracking norms and component contributions during alignment training (Section 4.1). We then test its information-preserving role through input replacement in pretrained models (Section 4.2). Finally, we examine the effects of mitigating norm inflation on representations and task performance by comparing training with and without norm regularization (Section 4.3).

4.1 Emergence of Norm Inflation during Modality Alignment

We compare norm trajectories and component contributions during alignment training under Baseline and Reduced initial norm.

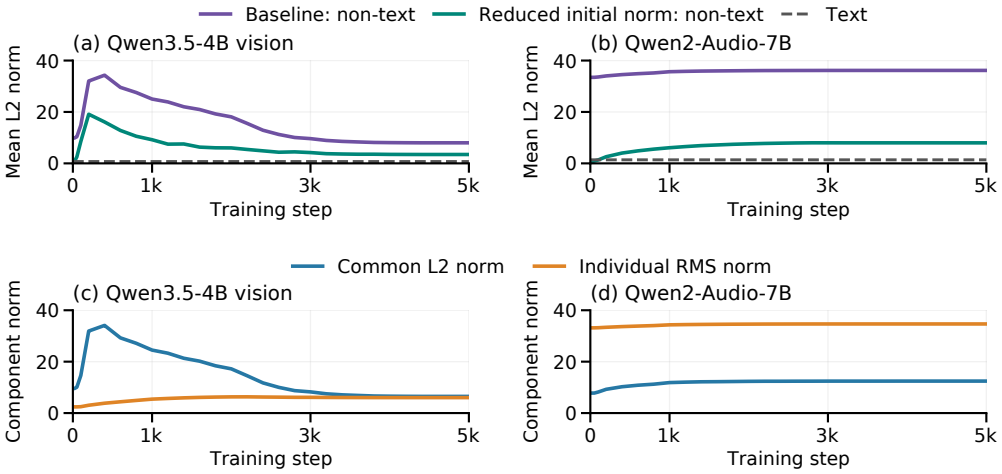


Figure 2: Input norms during alignment. (a–b) Mean L2 norms under both initializations and the text reference. (c–d) Common-component L2 and individual-component RMS norms for Baseline.

Inflation from small initial norms. Non-text norms exceed text norms after alignment even when initialized below them. Figure 2 shows input trajectories; Appendix D covers all recorded depths. The mean input norm of Qwen3.5-4B vision Baseline rises from 9.85 to 34.30 at step 400, then falls to 7.99. Under Reduced initial norm, it rises from 0.66 to 19.11 at step 200 and ends at 3.45, above the text norm of 0.71. Qwen2-Audio input norms rise from 33.47 to 36.16 under Baseline and from 0.72 to 7.98 under Reduced initial norm, exceeding the text norm of 1.38.

Component contributions to early inflation. The common component contributes substantially to early inflation, with its share varying by modality and initialization. Over steps 0–200, it accounts for 99.6% and 99.7% of the increase in mean squared input norm under vision Baseline and Reduced initial norm, respectively, versus 63.4% and 44.1% in audio. Later trajectories differ in Figure 2(c–d). Over vision Baseline steps 400–5,000, the common-component L2 norm falls from 34.10 to 6.48 while the individual-component RMS norm rises from 3.83 to 6.01. Variation across examples and token positions therefore becomes relatively more prominent. Audio Baseline's common-component norm instead rises from 9.24 at step 200 to 12.44 at step 5,000 and remains large.

Interchangeable inputs early in training. Replacing an image's input directions with those of another image tests how much early NTP improvement depends on the particular image. At vision Baseline step 100, this replacement barely changes loss, from 4.315 to 4.332, whereas Random at the same norm raises it to 5.459. By the end of training, Original and Other input losses diverge to 2.870 and 4.946; Appendix D.1 reports all measured steps. Early input interchangeability, together with the increasing common-component norm, suggests that a shared projector adjustment makes encoder and decoder spaces compatible and contributes to inflation. The later divergence indicates that image-specific representations become more important as alignment proceeds.

4.2 The Role of Norm Inflation in Information Preservation

A large norm can help preserve a decoder-compatible input across subsequent blocks by reducing the relative influence of updates on the pre-norm residual (Li et al., 2026). We test this role by replacing later non-text tokens with inputs at the same positions. The norm and direction controls in Section 3.2, together with Original representations, distinguish preserved inputs from subsequent transformations.

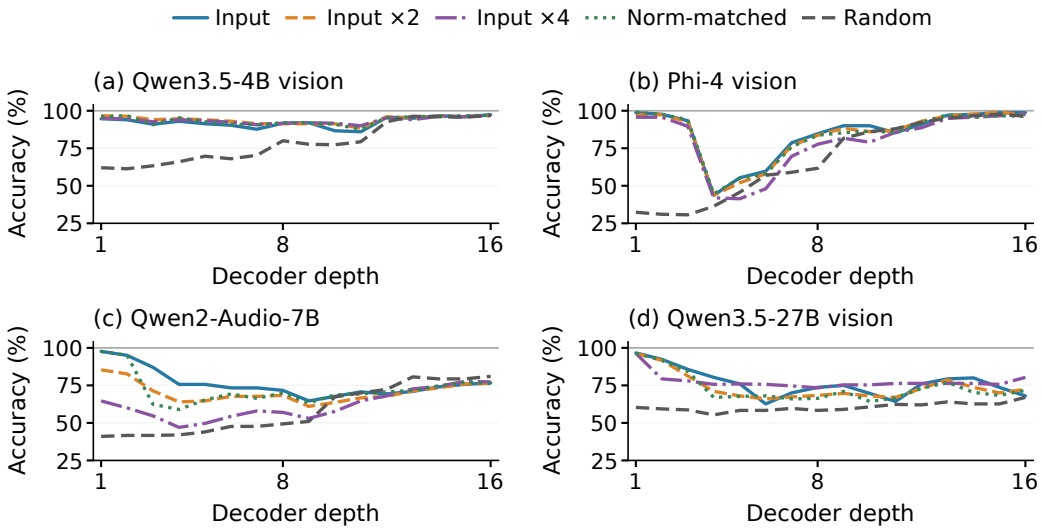


Figure 3: Accuracy after input replacement at different scales in pretrained models. Conditions follow Figure 1(b).

How well inputs replace later tokens. All four models in Figure 3 have early depths where Input retains at least 90% of originally correct answers. Accuracy is 90.3–94.7% at depths 1–6 in Qwen3.5-4B vision and 93.3–98.7% at depths 1–3 in Phi-4 vision. Input directions also outperform random directions at matched norms: at depth 3, Norm-matched and Random score 90.7% and 63.3% in 4B, and 94.0% and 30.7% in Phi-4. Encoder–projector inputs thus substitute for later representations and outperform random directions, supporting their preservation within the decoder. Appendix E.7 compares inputs conveying the same information as text.

Model-dependent effects of input scaling. Scaling inputs at a fixed direction can either improve or impair accuracy after replacement. At depth 3 of Qwen2-Audio in Figure 3(c), accuracy falls from 87.0% for Input to 71.3% and 54.7% for $\times 2$ and $\times 4$, even though Original tokens remain close to their inputs in direction (cosine similarity 0.962). At depth 6 of 27B in panel (d), $\times 4$ instead raises accuracy from 62.7% to 75.7%. Pre-norm normalization is approximately invariant to positive scaling, so enlarging the input increases its residual contribution relative to the first update; later blocks process the altered residual. Increasing this contribution from the input direction helps prediction in 27B and hinders it in Qwen2-Audio at the tested depths.

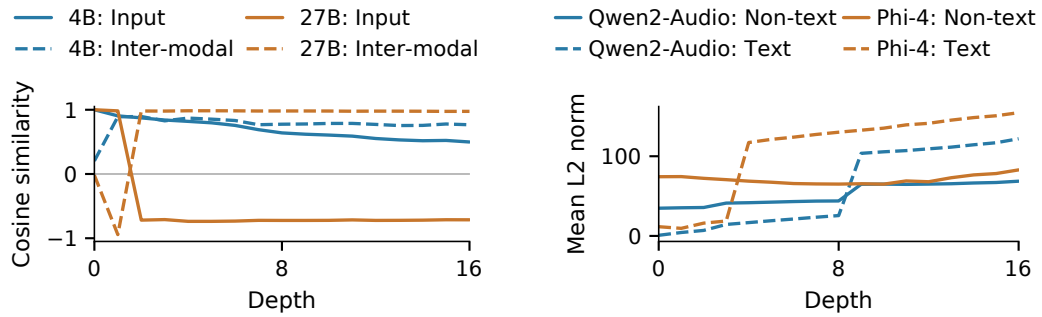


Figure 4: Original representations. Left: Qwen3.5-4B and 27B cosine similarity to inputs (solid) and between modality means (dotted). Right: mean non-text and text L2 norms in Phi-4 vision and Qwen2-Audio.

Lower accuracy after input replacement despite norm inflation. Despite norm inflation, accuracy after replacing later non-text tokens can decline as representations change within the decoder. In Qwen3.5-27B, Input accuracy falls from 96.7% at depth 1 to 62.7% at depth 6 while non-text norms exceed text norms. Input cosine similarity drops from 0.983 to -0.718 over depths 1–2 in Figure 4 (left), showing that these large-norm representations turn away from their inputs. Input direction still contributes at depth 6, where Norm-matched scores 68.0% versus Random's 58.3%. In Phi-4, the decline coincides with a reversal in relative modality norms. Over depths 3–4 in Figure 4 (right), text norms rise from 18.7 to 117.2 while non-text norms remain near 70 (70.6 to 68.7); Input accuracy falls from 93.3% to 43.7%. Qwen2-Audio's Input accuracy falls from 95.0% to 75.7% over depths 2–4, before the text norm surge. At depth 10, Input, Norm-matched, and Random all score 67–69%, compared with Original's 100%. Original's advantage over all three replacements shows that subsequent transformations contribute to prediction even where input direction offers little advantage over Random. Appendix E provides full trajectories and additional models.

Norms and information preservation. At early depths, inputs can replace later non-text tokens and outperform random directions at the same norm, supporting preservation of the input representations themselves. Changing norms at a fixed input direction changes accuracy, showing that the relative influence of the residual input and subsequent updates matters for prediction. Where accuracy after replacement declines, Original's advantage over Input shows that the decoder's transformations also contribute to prediction. Norm inflation can support input preservation during these transformations by reducing the relative influence of updates on the residual.

4.3 Effects of Mitigating Norm Inflation

We examine how mitigating norm inflation changes task performance and the balance between preservation and transformation in Qwen3.5-4B audio alignment. After reporting task results, we interpret the differences through training norms, Original representations, and input replacement.

Task performance. In Table 1, TAN_e improves transcription, question answering, and audio understanding over Baseline. For example, LibriMix WER falls from 9.66% to 7.70%, and MMAU accuracy rises from 59.1% to 69.0% (+9.9 percentage points). Initialization only captures much of the transcription and QA gains but scores 59.8% on MMAU. At the same initialization and learning rate, TAN_e gains another 9.2 points (95% confidence interval: [6.4, 12.0]). Input-only TAN_0 scores 54.0%, below Baseline. Regularized depths thus matter beyond initial norm reduction. Appendix F reports all benchmark results.

Table 1: Task performance after Qwen3.5-4B audio alignment. LS-clean denotes LibriSpeech test-clean.

Condition	LS-clean WER ↓	LibriMix WER ↓	SpokenSQuAD QA F1 ↑	MMAU Acc. ↑	MMAU-Pro Acc. ↑
Baseline	3.26	9.66	78.47	59.10	34.01
TAN_0	3.07	8.66	85.86	54.00	33.90
TAN_e	2.81	**7.70**	**87.32**	**69.00**	**35.86**
Initialization only	**2.76**	8.11	86.53	59.80	33.66

TAN_e mitigates internal norm inflation. Figure 5 shows that both TAN conditions mitigate input norm inflation during projector training. The norm rises again in TAN_0's first block, whereas TAN_e keeps norms close to text through regularized depth 11. For example, at short-stage step 1,668, first-block non-text and text norms are 33.93 and 2.93 for TAN_0, versus 4.78 and 2.92 for TAN_e. Appendix E.3 covers all depths and recorded steps.

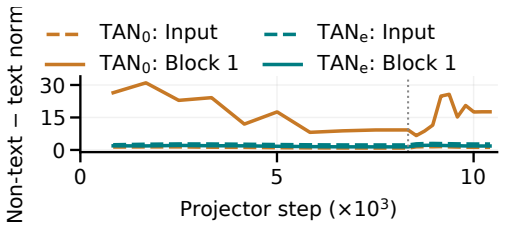


Figure 5: Mean non-text minus text norm. The dashed line marks the start of long-stage training.

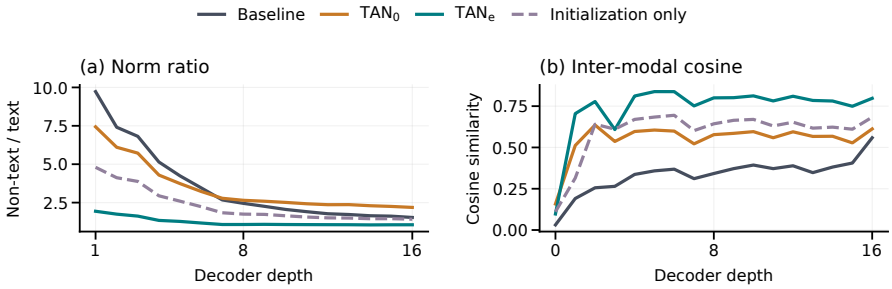


Figure 6: Original representations after TAN. (a) Mean of per-example ratios between mean non-text and text norms. (b) Cosine similarity between the two modality mean vectors.

Changes in direction relative to the input. Table 2 shows that after LoRA, Original non-text tokens depart from their inputs progressively under Baseline and earlier under both TAN conditions. At depths 3 and 12, input cosine similarity is 0.941 and 0.569 under Baseline, versus 0.505 and 0.184 under TAN_e. By depth 32, similarity to their respective inputs is low: 0.120 and 0.091. Under TAN_e, this earlier change accompanies small norms (Figure 6(a)): the depth-3 non-text norm is 5.35, or 1.62 times the text norm. Under TAN_0, first-block input cosine similarity falls to 0.431 as the norm rises from 1.88 to 18.13. Both conditions transform inputs early, but TAN_e does so with internal norm inflation mitigated.

Table 2: Input cosine similarity across decoder depth. †: ablation without norm regularization.

Depth	1	3	6	12	20	28	32
Baseline	0.976	0.941	0.841	0.569	0.331	0.170	0.120
TAN_0	0.431	0.341	0.228	0.132	0.092	0.058	0.104
TAN_e	0.681	0.505	0.300	0.184	0.127	0.087	0.091
Initialization only†	0.844	0.690	0.546	0.343	0.204	0.120	0.121

Higher inter-modal similarity under TAN_e. With norm inflation mitigated, TAN_e achieves higher inter-modal cosine similarity and reaches high values earlier. In Figure 6(b), first-block similarity is 0.705 versus Baseline's 0.191. At depth 6, TAN_e reaches 0.838, versus 0.368 under Baseline, 0.599 under TAN_0, and 0.695 under Initialization only. Internal norm regularization thus brings modality mean directions closer than initial norm reduction alone.

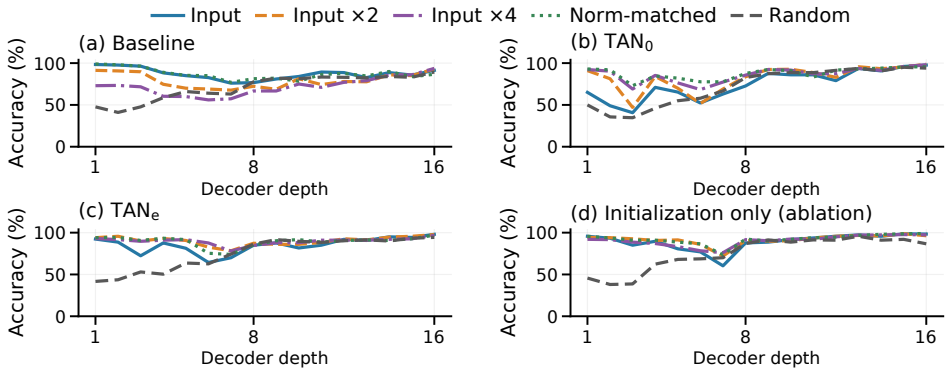


Figure 7: Accuracy after input replacement at different scales after TAN.

Reduced input preservation under norm regularization. Figure 7 shows lower Input accuracy under TAN_e than Baseline throughout depths 1–6: 72.3% versus 96.3% at depth 3. Together with earlier directional changes, this indicates reduced preservation of unchanged inputs. Rescaling reveals a remaining contribution from input directions: TAN_e's Norm-matched scores 75.3–94.3% over depths 1–6, exceeding Random at every depth (90.7% versus 53.0% at depth 3). Under TAN_0, Norm-matched scores 73.0–93.0%, above Input's 40.7–71.0%. Norm regularization thus reduces preservation of unchanged inputs, with some information preservation evident after norm adjustment.

Representation changes after mitigating norm inflation. TAN_e reduces input preservation and improves audio–text similarity and task performance relative to Baseline. Some preservation remains after norm adjustment. Component analysis and cross-input replacement in Section 4.1 suggest that shared adjustments make encoder and decoder spaces compatible and contribute to early inflation. In audio, the common-component norm remains large late in training. Large norms help preserve decoder-compatible inputs across later blocks by reducing the relative influence of updates on the residual. The replacement controls in Section 4.2 support this role. When the decoder largely retains non-text inputs while adapting representations for NTP, further transformation with text may be limited. We interpret TAN_e's earlier departure from input directions and higher audio–text similarity as earlier joint representation formation than under Baseline. The gains over the matched-initialization ablation suggest that joint audio–text representations help the pretrained decoder apply its capabilities to audio tasks.

5 Conclusion

We explain norm inflation during modality alignment through its role in preserving decoder-compatible inputs. Inflation emerges even from reduced initial norms. Component analysis and cross-input replacement suggest that shared adjustments make encoder and decoder spaces compatible and contribute to early inflation. In pretrained models, input tokens can replace later non-text representations while sustaining predictions. Norm and direction controls support a role for large norms in preserving non-text inputs. The study is limited to pre-norm decoders, with norm mitigation evaluated in Qwen3.5-4B audio alignment. Regularizing norms through the early decoder with TAN_e reduces input preservation and improves inter-modal similarity and task performance relative to Baseline. Earlier directional changes and the preservation that remains after norm adjustment suggest earlier joint audio–text representation formation. The observed task gains suggest that joint representations help the pretrained decoder apply its capabilities to audio tasks.

AI USE STATEMENT

Generative AI provided extensive support for experiments and manuscript preparation. The authors developed the research ideas and interpreted the results; AI helped specify protocols for the experimental questions set by the authors. The authors selected and revised these protocols. AI coding agents wrote nearly all of the current experimental code and inspected code and ran tests under the authors' instructions. This included implementing procedural visual questions and restructuring existing audio recordings and transcripts. The authors developed the manuscript's argument and structure and directly wrote and revised its sentences. AI supported writing and translation based on the authors' text and sentence-level revision instructions. The references were assembled from materials collected by the authors; AI was used to cross-check bibliographic information and publication status. AI also assisted with figures and tables. The authors are responsible for the final manuscript, research methods, code, and reported results.

ETHICS STATEMENT

This study analyzes internal MLLM representations using existing datasets and controlled questions. We recruited no new participants, conducted no human-subject experiments, and collected no new personal data. Controlled visual questions were generated procedurally, and controlled audio questions were constructed from existing recordings and transcripts. Use and redistribution of third-party datasets and models are subject to their original licenses and access conditions.

REPRODUCIBILITY STATEMENT

Section 3 defines the training and replacement procedures. Appendix B provides training data, optimization, and initialization settings; Appendix C details evaluation cohorts, aggregation, and replacement validation. Appendix F specifies benchmark generation and scoring settings and metric-specific denominators. We provide training and evaluation code, experimental configurations, evaluation-cohort construction procedures, and the result files underlying the reported values in a public repository. The repository records dependencies, model revisions, data sources, and commands for each experiment, together with the scope and numerical tolerances of sampled reruns validating the stored results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. ALLaVA: Harnessing GPT4V-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. URL https://arxiv.org/abs/2402.11684.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628. URL https://aclanthology.org/2026.tacl-1.18/.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. LibriMix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*, 2020. URL https://arxiv.org/abs/2005.11262.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27092–27112, 2023. doi: 10.52202/075280-1179. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56332d41d55ad7ad8024aac625881be7-Abstract-Datasets_and_Benchmarks.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026. URL https://arxiv.org/abs/2604.10905.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Han_OneLLM_One_Framework_to_Align_All_Modalities_with_Language_CVPR_2024_paper.html.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jiang_Hallucination_Augmented_Contrastive_Learning_for_Multimodal_Large_Language_Model_CVPR_2024_paper.html.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. MMAU-Pro: A

challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22688–22697, 2026. doi: 10.1609/aaai.v40i27.39430. URL https://ojs.aaai.org/index.php/AAAI/article/view/39430.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*, pp. 3459–3463, 2018. doi: 10.21437/Interspeech.2018-1714. URL https://www.isca-archive.org/interspeech_2018/lee18d_interspeech.html.

Bozhou Li, Xinda Xue, Sihan Yang, Yang Shi, Xinlong Chen, Yushuo Guan, Yuanxing Zhang, and Wentao Zhang. The unseen bias: How norm discrepancy in pre-norm MLLMs leads to visual information loss. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/03de675250598f76340d459c8bbf75d5-Abstract-Conference.html.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022. doi: 10.52202/068431-1280. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. URL https://arxiv.org/abs/2508.11737.

Ahmed Masry, Juan A. Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, et al. AlignVLM: Bridging vision and language latent spaces for multimodal document understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 161722–161748, 2025. doi: 10.52202/085713-4875. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/d695f1d2b37df1310f3b9d2bbc7e0a2f-Abstract-Conference.html.

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*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://www.danielpovey.com/files/2015_icassp_librispeech.pdf.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya

Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d36f208919582785db965fe648b9fe59-Abstract-Conference.html.

Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen LLMs to multimodal inputs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130848–130886, 2024. doi: 10.52202/079017-4159. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ec3c79dc0c2b85532cfd1012a4aaa923-Abstract-Conference.html.

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 *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *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. 4297–4316, 2025. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, 2025. doi: 10.18653/v1/2025.emnlp-main.262. URL https://aclanthology.org/2025.emnlp-main.262/.

Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Yuanxing Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Wentao Zhang, and Feng Zhao. SEA: Supervised embedding alignment for token-level visual-textual integration in MLLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1058–1070, 2025. doi: 10.18653/v1/2025.emnlp-main.55. URL https://aclanthology.org/2025.emnlp-main.55/.

A Detailed Related Work

We review modality alignment in MLLMs through their connecting architectures and training objectives, followed by studies of the modality gap within the decoder and explanations and interventions concerning norm discrepancy.

A.1 Modality alignment in MLLMs

MLLMs combining pretrained encoders and LLMs have developed different connecting architectures and training procedures. Flamingo introduces cross-attention between frozen vision and language models to convey encoder information within the decoder (Alayrac et al., 2022). BLIP-2 connects visual encoder outputs to an LLM through a Q-Former with learnable queries (Li et al., 2023). LLaVA uses projected features as LLM inputs and learns instruction following from image-related instruction–response data (Liu et al., 2023; 2024). SALMONN connects speech and general audio encoder outputs to an LLM to jointly learn transcription and audio understanding (Tang et al., 2024). These models learn text generation conditioned on non-text inputs while varying where and how encoder information enters the LLM.

Other methods share the architecture connecting modalities to the LLM or directly use its text embeddings. OneLLM aligns eight modalities, including images, audio, video, and point clouds, with language through a shared encoder and a universal projection module (Han et al., 2024). The module combines outputs from multiple projections with weights determined by input features, progressively extending alignment to additional modalities. AlignVLM derives weights over LLM vocabulary embeddings from visual features and uses their weighted sum as LLM input (Masry et al., 2025). By representing visual features as combinations of pretrained text embeddings, it improves document understanding and performance with limited training data.

Training objectives can also directly learn correspondences between modalities. HACL analyzes the gap between image and text representations and the overlap between correct descriptions and hallucinated text, then proposes contrastive learning with hallucinated text as hard negatives (Jiang et al., 2024). Training brings images closer to correct descriptions and farther from hallucinated text, reducing hallucination on evaluation benchmarks. SEA trains the connecting module by adding token-level visual–text embedding alignment to image-level supervision (Yin et al., 2025). Both approaches supply a learning signal for semantic correspondence between representations alongside prediction of generated responses.

A.2 Modality gap and functional use of representations

In contrastive models, Liang et al. (2022) showed that encoder initialization and contrastive optimization contribute to the formation and persistence of the modality gap. In MLLMs with frozen LLMs, Shukor & Cord (2024) compared representations obtained by averaging each input's tokens within each modality. Within the decoder, cosine similarity remained high among mean representations of the same modality and low between non-text and text, indicating concentration in different directions. They also reported larger norms and smaller directional changes across depths for non-text tokens than for text tokens. During single-task training of a mapping module connected to Vicuna-v1.5, non-text norms increased across epochs.

Xiang et al. (2025) separately fed speech and its transcript to an MLLM and compared the two modality mean representations at the same decoder depth. Both cosine similarity and Euclidean distance increased with depth. For vectors u, v, $\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \cos(u, v)$, so changes in norms can increase distance even as directions become closer. The results show that directional alignment and increasing distance between representations can coexist within the decoder.

Studies have also examined the implications of representation differences for task performance. Across training checkpoints and LLaVA variants, Shukor & Cord (2024) found that higher cosine similarity between the input non-text mean representation and text mean representations at later depths tended to accompany higher benchmark scores. They also identified settings where skipping FFN computation for selected visual tokens caused little degradation, testing how much of the subsequent transformation was needed for the task. Xiang et al. (2025) intervened in trained

models by matching the direction or norm of selected non-text tokens to corresponding text tokens. On VoiceBench sd-qa, they compared eight settings spanning two decoders, two fine-tuning methods, and two token-selection scopes. Matching directions maintained or improved QA accuracy in six settings, whereas matching norms reduced it in seven. Thus, the effect on correct predictions depends on which property of the representation is changed.

A.3 Norm discrepancy in pre-norm decoders

Li et al. (2026) analyzed the relationship between large visual-token norms and small directional changes in pre-norm MLLMs, interpreting the limited transformation of large-norm representations as delaying inter-modal integration. They proposed a learnable LayerNorm after the visual projector, initialized its gain from text norms, and corrected for reduced gradients. The LayerNorm remains in the forward pass during both training and inference. Their study includes analyses across depths and training checkpoints and reports performance improvements after norm control. Our TAN uses a training loss referenced to text norm statistics at the same depth and extends regularization from the input through the early decoder.

B Training Details

We provide the training data, optimization, initialization, and TAN settings for Sections 3.1 and 3.3. Formation experiments track changes during projector training; TAN comparisons include projector training followed by decoder LoRA.

B.1 Data and optimization

Vision formation experiments use 60,000 image–text examples from DataComp (Gadre et al., 2023), ALLaVA (Chen et al., 2024), LLaVA alignment, LVIS-Instruct4V, ShareGPT4Video, Vision-FLAN, and ScienceQA. Audio data comprise captions from AudioCaps, Clotho, MusicCaps, and WavCaps; transcripts from GigaSpeech, LibriSpeech, and LibriMix; and question answering and transcription examples from SpokenSQuAD. Audio formation experiments use 60,000 examples. Both initializations of a model use the same training data and evaluation probes.

For TAN, the projector short stage uses 200,000 examples. The long stage uses 40,000 new examples spanning a wider range of input lengths and 10,000 examples from the short stage. Decoder LoRA uses 15,000 examples from these data. Short-stage and decoder sequences contain at most 2,048 tokens; long-stage sequences contain at most 7,639 tokens. The new long-stage data also include short inputs. Table 3 gives the source composition. The projector is a biased linear layer mapping the audio encoder's 1,280-dimensional outputs to 2,560-dimensional decoder inputs. The encoder is Qwen2.5-Omni-3B and the decoder is Qwen3.5-4B. Evaluation data used during training and the main LibriSpeech, LibriMix, SpokenSQuAD, and MMAU evaluation sets do not overlap with training data by recorded audio path. Table 4 summarizes optimization settings.

Table 3: Training examples by source in the TAN comparison. The 10,000 reused long-stage examples and 15,000 decoder-stage examples come from earlier projector training data.

Source	Projector short	Projector long	Decoder LoRA
AudioCaps	33,014	8,038	2,100
Clotho	2,893	600	2,100
GigaSpeech	35,112	9,635	900
LibriMix	29,674	5,800	900
LibriSpeech clean	30,099	7,000	900
LibriSpeech other	30,306	7,079	900
MusicCaps	2,581	600	2,100
SpokenSQuAD QA	12,554	3,406	2,100
SpokenSQuAD STT	14,809	5,486	900
WavCaps	8,958	2,356	2,100
Total	200,000	50,000	15,000

Table 4: Training settings. Both experiments use BF16 and seed 122. Learning rates are the maxima for each stage.

Setting	Formation experiments	TAN comparison
Trainable components	Projector	Projector, then decoder LoRA
Global batch	12	24
Training duration	5,000 steps	1 epoch per stage
Default-initialization projector LR	10^{-4}	10^{-4}
Reduced-norm-initialization projector LR	5×10^{-5}	5×10^{-5}
Decoder LoRA LR	Not applicable	5×10^{-6}
Optimizer	AdamW	AdamW
(β_1, β_2) / ϵ	$(0.9, 0.95)$ / 10^{-8}	$(0.9, 0.95)$ / 10^{-8}
Weight decay / gradient clip	0.01 / 1.0	0.01 / 1.0
Schedule / warmup fraction	Cosine / 0.03	Cosine / 0.03
LoRA rank / alpha / dropout	Not applicable	8 / 8 / 0

Per-device microbatch size for TAN is 4 in the short stage and 1 in the long and decoder stages. Gradient accumulation yields a global batch size of 24. LoRA is applied to decoder attention, linear-attention, and FFN projections; biases remain frozen. The targets are the self-attention q/k/v/o projections, linear-attention in_proj_qkv/z/b/a and out_proj, and FFN gate/up/down projections.

Formation measurements are taken at steps 0, 1, 2, 5, 10, 20, 50, 100, 200, 400, and every 200 steps from 600 through 5,000: 33 checkpoints per run and 132 across two models and two initializations. TAN_0 and TAN_e trajectories are measured at 10 checkpoints in each of the short and long stages using the same 256 items. The decoder remains frozen at its pretrained state during these measurements. One epoch ends at step 8,334 for projector short training, 2,084 for projector long training, and 625 for decoder LoRA. The final 300-item analyses and task evaluations use the completed LoRA checkpoint.

B.2 Initialization and norm regularization

Formation initialization. Formation experiments reinitialize trainable layers of the existing projector architecture with `reset_parameters`. The Qwen3.5-4B vision merger comprises a 1,024-dimensional LayerNorm, concatenation of adjacent 2×2 visual tokens, a $4,096 \to 4,096$ linear layer, GELU, and a $4,096 \to 2,560$ linear layer. The Qwen2-Audio projector is a $1,280 \to 4,096$ linear layer; linear layers in both architectures include biases. Both conditions use the same initial weights saved with seed 122. Reduced initial norm applies a single positive scale to the final linear layer's weight and bias. The scale is the ratio of the median valid text norm to the median initial non-text norm collected from the first 256 training examples: 0.0665934 for vision and 0.0216181 for audio.

TAN initialization. TAN and Initialization only in Qwen3.5-4B audio alignment share the same reduced-norm initialization. The median norm of 16,384 audio-feature tokens passed through the new linear projector is 21.650309, and the vocabulary-embedding median is 0.651979. Their ratio, $\alpha = 0.0301141$, scales both the projector weight and bias. Comparison with Initialization only measures the effect of adding norm regularization from the same initial state.

Text norm reference statistics. The input reference is computed from the L2 norms of all 248,320 vocabulary embeddings, including special and reserved tokens. The BF16 execution values have mean 0.655775 and standard deviation 0.067300. Within the decoder, references use valid text positions from the same forward pass, including the training prompt and response. Given median norm m and median absolute deviation d_{MAD}, the reference mean and standard deviation use values satisfying $|r - m| \leq 3.5 \times 1.4826 d_{\text{MAD}}$. Gradients do not flow through text references. Representation analyses after training measure prompt text before response generation, excluding non-text and special positions.

Microbatch statistics. Non-text and decoder text norm statistics are computed per device microbatch, weighting valid tokens equally. Longer examples therefore contribute more tokens. Standard deviations use the population definition. The loss in Equation (2) is computed separately for each

microbatch during gradient accumulation. The standard-deviation term applies when each side contains at least two tokens; a depth's regularization term is omitted if its text reference is empty.

Regularization coefficient and depths. The coefficient $\lambda = 0.1$ in Equation (3) is the fixed weight used in earlier alignment-loss comparisons. TAN_0 applies it at the input, and TAN_e applies $0.1/12$ at each depth from 0 to 11. We set $e = 11$ from an earlier Qwen3.5-4B vision Random replacement curve, using the point beyond which replacement had smaller effects at later depths. With the same λ, the two conditions jointly compare the extent of regularization and its strength at each depth.

C Evaluation Cohorts and Metrics

We describe the cohorts, aggregation, and measurement checks used for representation and replacement analyses. Appendix F covers task-specific scoring of generated responses.

C.1 Fixed cohorts and replacement checks

The four vision sources are procedurally generated questions about colors, shapes, and counts; MMBench dev; MMMU validation; and ScienceQA test. Audio sources are transcript-selection questions from LibriSpeech, LibriMix, and SpokenSQuAD; MMAU test-mini; multiple-choice MMAU-Pro; and VoiceBench OpenBookQA and MMSU (Chen et al., 2026). Existing benchmark questions, options, and answers are retained. Main-text training trajectories use 256 items, 64 per source, selected independently of correctness. Appendix D.2 describes measurement cohorts for the additional training models.

Replacement analyses for the six pretrained model–modality conditions each select 75 originally correct items per source. TAN comparisons select 75 per source from items answered correctly by all four training conditions. Each 300-item cohort, fixed with seed 122, is used across replacement conditions and depths. Cohorts differ between pretrained models; the four TAN conditions share one cohort. Original representation statistics for Phi-4 audio and Qwen2.5-Omni-3B audio use a separate fixed set of 256 probes. Input scales of 2 and 4 are exploratory comparisons chosen with reference to a separate depth-3 exploration.

Depth ℓ denotes the output boundary of block ℓ; a forward pass replaces only non-text positions at that boundary. Replaying the identical hidden states produced a maximum difference of 0 in option logits relative to Original. The maximum relative error in Random's target norm was 2.91×10^{-4} for pretrained comparisons and 3.18×10^{-4} for TAN. When combining Norm-matched results, we verified agreement of Input and Original logits for the same item and depth. Replacing non-text tokens after the final block may leave the already computed response position unaffected, so high accuracy at this boundary is not used as evidence of information preservation.

C.2 Aggregation and uncertainty

Norms and input cosine similarities are first averaged over valid tokens within each example and then equally across examples. A per-example norm ratio divides mean non-text norm by mean text norm; these ratios are averaged across examples. Inter-modal cosine similarity is computed between the non-text and text mean vectors within each example. Padding is excluded, and examples with more tokens receive the same weight as others.

Benchmark 95% confidence intervals use 5,000 bootstrap replicates (seed 122) over clusters connecting items that share a recording. Differences between conditions in supplementary replacement experiments use paired bootstrap resampling of the same clusters. Each interval describes sampling uncertainty over the evaluation items for that task and depth.

D Formation Trajectories and Component Contributions

We present training trajectories for Qwen3.5-4B and 9B vision, Phi-4 vision, and Qwen2-Audio-7B. For the two main-text models, we first describe input norms, component contributions over the first 200 steps, and changes across all recorded training steps and decoder depths. Appendix D.1

examines early cross-input compatibility; Appendix D.2 provides additional trajectories for 9B and Phi-4. Appendix E.1 compares models in their pretrained states.

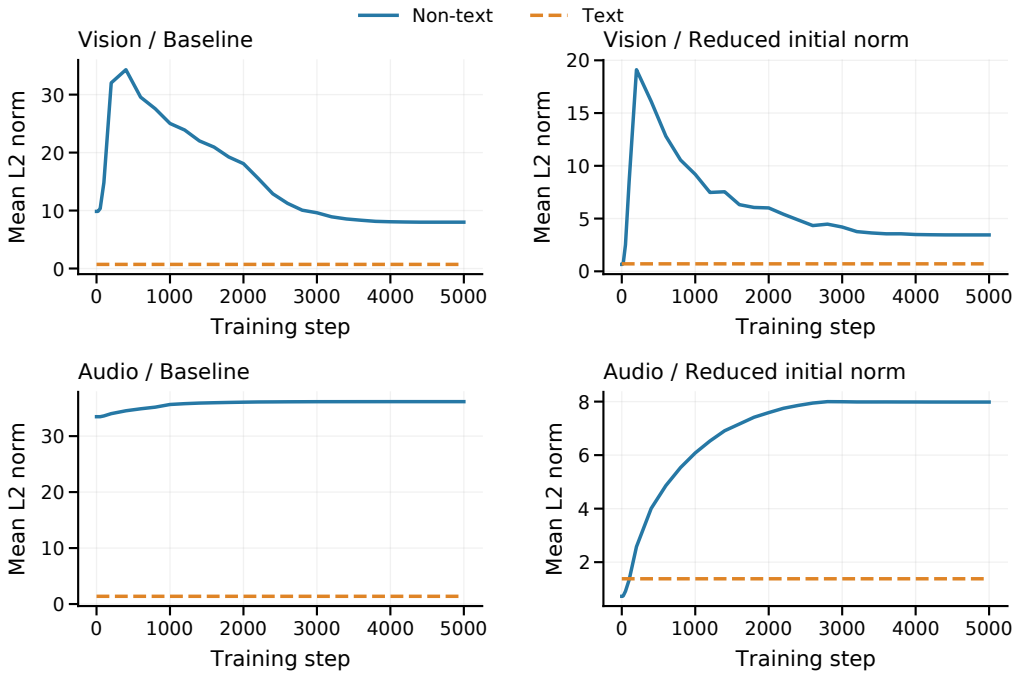


Figure 8: Mean input L2 norms for Baseline and Reduced initial norm. Top: Qwen3.5-4B vision. Bottom: Qwen2-Audio-7B.

Component decomposition and contributions. We use mean squared norms to decompose early inflation into two contributions. Let i index valid token positions within example x, with equal weights for tokens within an example and for examples across the cohort. For the common component $c = \mathbb{E}_x \mathbb{E}_{i|x} h_{x,i}$ and individual component $h_{x,i} - c$, the mean cross term is zero because $\mathbb{E}_{x,i}(h_{x,i} - c) = 0$. Thus,

$$E = \mathbb{E}_{x,i}\|h_{x,i}\|_2^2 = \underbrace{\|c\|_2^2}_{E_c} + \underbrace{\mathbb{E}_{x,i}\|h_{x,i} - c\|_2^2}_{E_i}. \tag{4}$$

For training interval $[t_0, t_1]$, define $\Delta E = E(t_1) - E(t_0)$, giving $\Delta E = \Delta E_c + \Delta E_i$. Where the mean squared norm increases ($\Delta E > 0$), the common and individual contributions are $100\Delta E_c/\Delta E$ and $100\Delta E_i/\Delta E$, respectively. Table 5 gives the increases over steps 0–200. For example, of vision Baseline's total increase of 930.07, the common component contributes 926.59 and the individual component 3.48. The common contribution, calculated before rounding, is 99.6%.

Relating component norms to contributions. Figure 2(c–d) plots the common-component L2 norm $\|c\|_2 = \sqrt{E_c}$ and the root mean square (RMS) individual-component norm $\sqrt{\mathbb{E}\|h - c\|_2^2} = \sqrt{E_i}$. These trajectories express component magnitudes in the same norm units. Contributions are calculated by squaring values at each step and taking each term's increase as a fraction of the total increase in Equation (4). In general, $\sqrt{E_c} + \sqrt{E_i} \neq \sqrt{E}$; $\sqrt{E}$ also differs from the arithmetic mean L2 norm $\mathbb{E}\|h\|_2$ in Figure 2(a–b). Component norms describe magnitudes at individual steps; contributions quantify their shares of the increase in mean squared norm over a training interval.

Decomposing the individual-component mean squared norm. With example mean $m_x = \mathbb{E}_{i|x} h_{x,i}$, the individual-component mean squared norm decomposes as

$$\mathbb{E}_{x,i}\|h_{x,i} - c\|_2^2 = \mathbb{E}_x\|m_x - c\|_2^2 + \mathbb{E}_{x,i}\|h_{x,i} - m_x\|_2^2. \tag{5}$$

The first term measures differences between example means; the second measures token differences within examples. Their sum E_i is the individual-component mean squared norm used in the main text and, together with E_c, constitutes the total mean squared norm.

Table 5: Increase in mean squared input norm over steps 0–200. $\Delta E = \Delta E_c + \Delta E_i$; contributions are calculated before rounding the displayed values.

Model	Initialization	ΔE	ΔE_c	ΔE_i	Common share (%)
Vision	Baseline	930.07	926.59	3.48	99.6
Vision	Reduced initial norm	365.61	364.39	1.22	99.7
Audio	Baseline	39.31	24.91	14.40	63.4
Audio	Reduced initial norm	6.91	3.05	3.86	44.1

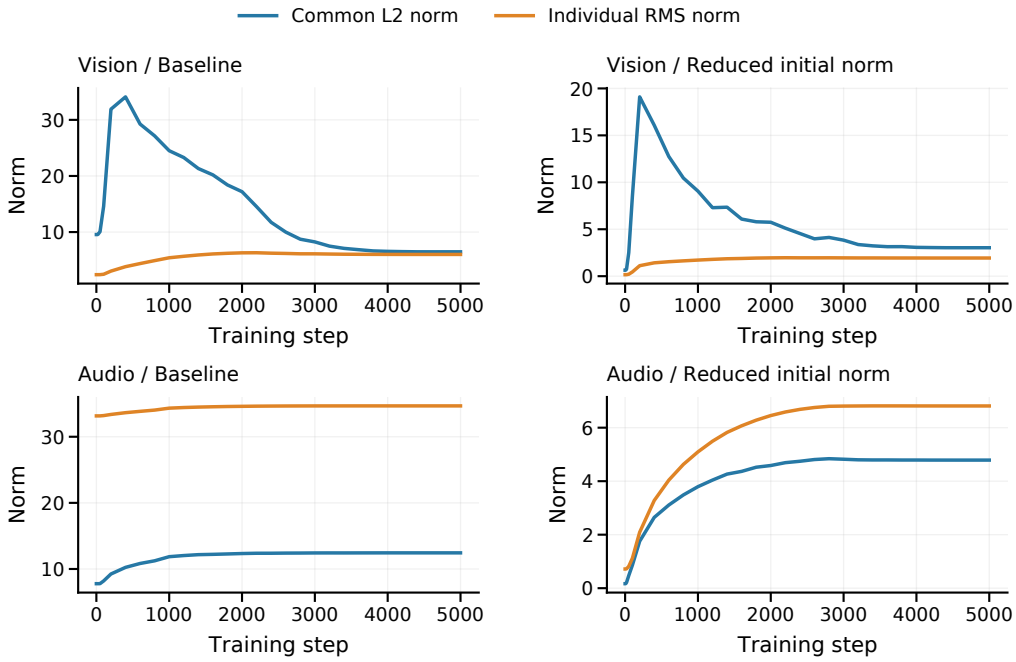


Figure 9: Full training trajectories of input common-component L2 norms and individual-component RMS norms under both initializations.

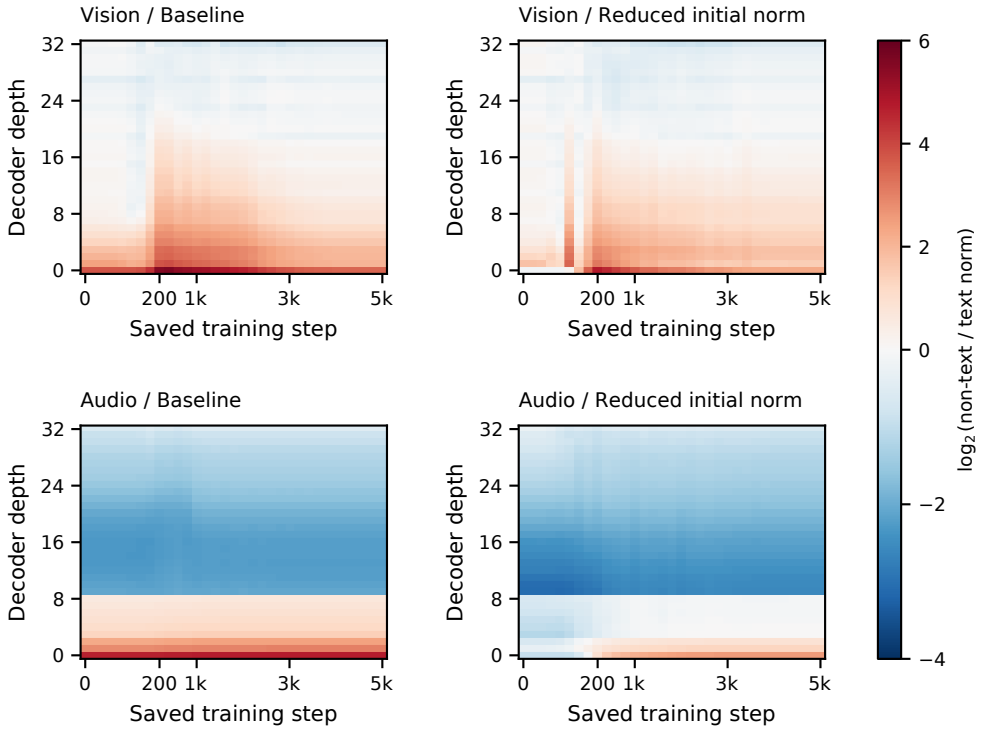


Figure 10: Mean norm ratios across all depths and checkpoints in the four formation runs. Each cell is the $\log_2$ ratio of mean non-text to text norms aggregated across examples. Zero denotes equal norms; one denotes a twofold ratio. All panels share a color scale. Horizontal cells correspond to 33 saved checkpoints, whose training-step intervals are unequal.

D.1 Input compatibility during early alignment

Cross-input replacement and NTP loss. To examine the interpretation of early norm increases as a common adjustment between representation spaces, we replaced image input tokens with another image's tokens at Qwen3.5-4B vision Baseline training checkpoints. Of 256 separate alignment-response items, 248 had another image with the same non-text token count. The other image was fixed by a hash of the item identifier; the original text, response, and positional information were retained. **Original** uses the item's own input. **Other input** uses another image's input directions with the original input's per-token norms. **Common** uses one direction at every non-text position, obtained by averaging non-text tokens within each item and then across 384 items: the 256 items above and 128 training examples. **Random** uses random directions. Common and Random also retain the original input's per-token norms. Responses are teacher-forced to obtain mean NTP loss over response tokens, weighting the 248 items equally. All conditions are compared at steps 0, 20, 50, 100, 200, 400, 1,000, and 5,000.

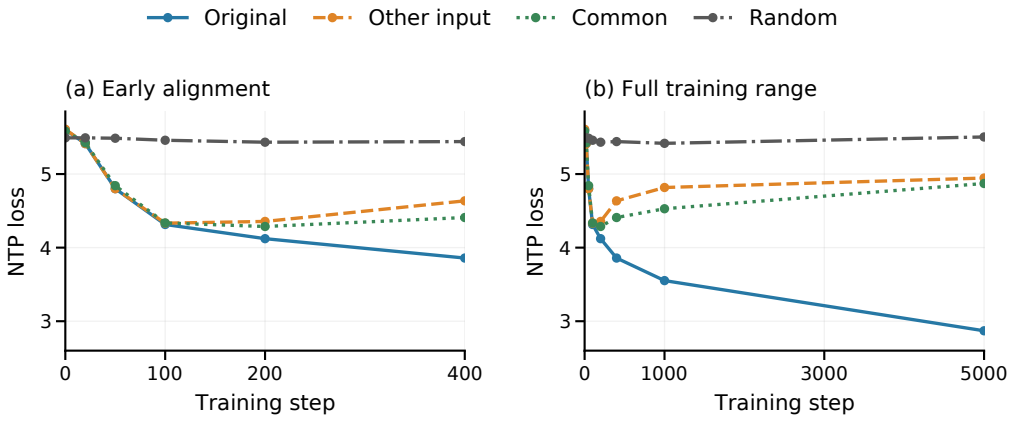


Figure 11: NTP loss by input direction in Qwen3.5-4B vision Baseline. (a) Early training. (b) All measurements through the end of training. Lines connect measured points.

Interchangeable inputs early in training. During early training in Figure 11, NTP loss decreases similarly for the original input, another input, and the common mean direction, while remaining higher for random directions. Original loss falls from 5.601 at step 0 to 4.315 at step 100; Other input loss falls from 5.609 to 4.332. At step 100, Common yields a loss of 4.335 and Random 5.459. The Original–Other input gap widens from step 200, reaching losses of 2.870 and 4.946 at the end of training. Table 6 provides all measurements. The similar early losses suggest that adjustments shared across inputs improve decoder prediction, with image-specific representations contributing increasingly as training proceeds. Together with the common-component norm increase in Section 4.1, these observations support interpreting early inflation as partly reflecting an adjustment that makes encoder and decoder representation spaces compatible.

Table 6: NTP loss under cross-input replacement. Δ NTP is the mean of Other input loss minus Original loss on the same items.

Step	Original	Other input	Common	Random	Δ NTP
0	5.601	5.609	5.582	5.492	0.008
20	5.414	5.413	5.425	5.492	-0.001
50	4.801	4.801	4.842	5.487	-0.001
100	4.315	4.332	4.335	5.459	0.018
200	4.122	4.356	4.287	5.433	0.234
400	3.858	4.636	4.409	5.441	0.777
1000	3.552	4.817	4.528	5.417	1.265
5000	2.870	4.946	4.872	5.504	2.076

D.2 Formation in additional vision models

Models and trajectory comparisons. We also trained projectors for Qwen3.5-9B and Phi-4 vision for 5,000 steps with frozen encoders and decoders. Learning rates, global batch size, and checkpoints follow the formation settings in Appendix B. Phi-4 Baseline and Reduced initial norm are measured over depths 0–32 using 256 fixed probes from the same four sources as the main text. The initial norm scale, estimated from training data, is 0.642704. Qwen3.5-9B Baseline input norms are measured on a separate set of 256 image–alignment-response probes and compared with 4B on the same probes. Each curve connects 33 measurements over steps 0–5,000.

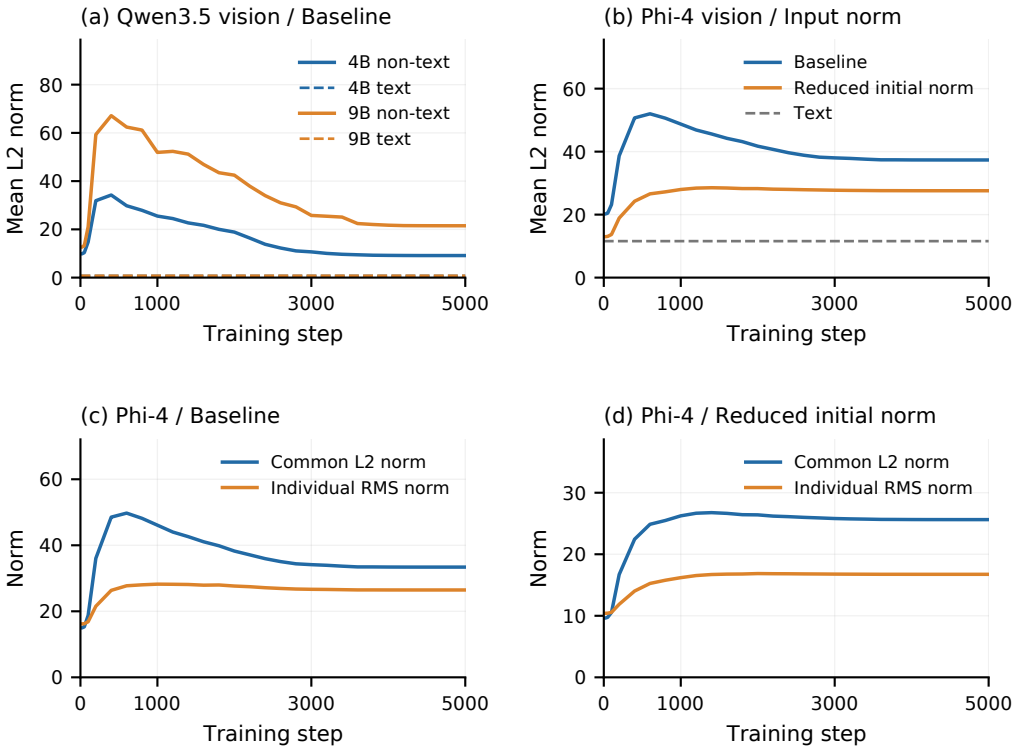


Figure 12: Full training trajectories in additional vision models. (a) Qwen3.5-4B and 9B Baseline input norms on the same probes. (b) Phi-4 input norms under both initializations. (c–d) Phi-4 input common-component L2 norms and individual-component RMS norms.

Inflation and contraction in additional models. In Figure 12(a), Qwen3.5-9B input norms rise from 12.64 to 67.14 at step 400, then fall to 21.49 at step 5,000. On the same probes, 4B input norms change from 9.85 to 34.24 and then 9.14. Both models exhibit inflation followed by contraction; 9B has a larger increase and ends above its initial norm. Phi-4 Baseline input norms likewise rise from 20.20 to 52.01 at step 600, then fall to 37.35. Under Reduced initial norm, they rise from 12.99 to 28.54 at step 1,400 and remain near that level, ending at 27.58. The text norm is 11.58, so the non-text/text norm difference increases during training even from the smaller initialization.

Common contributions and differing contraction patterns. Over Phi-4's first 200 steps, the common component accounts for 84.2% of the increase in mean squared norm under Baseline and 85.1% under Reduced initial norm. The substantial common contribution to early inflation therefore extends to this model. Over Baseline steps 600–5,000, the common-component L2 norm falls from 49.75 to 33.39, and the individual-component RMS norm decreases slightly from 27.73 to 26.46. Qwen3.5-4B also shows a decrease in common-component norm during contraction, accompanied by an increase in individual-component RMS norm. Figure 13 shows norm relationships across all depths and recorded steps under both Phi-4 initializations.

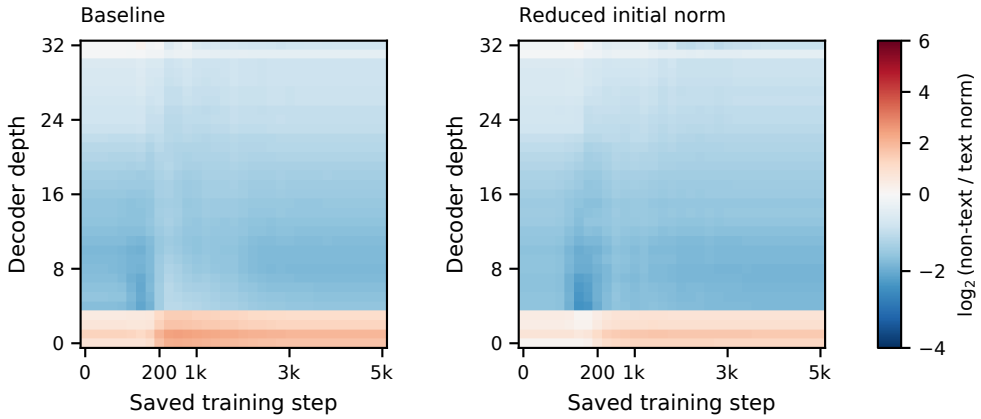


Figure 13: Mean norm ratios across all depths and training checkpoints in Phi-4 vision. Colors give the $\log_2$ ratio of mean non-text to text norms on the same scale as Figure 10. Horizontal cells correspond to saved checkpoints.

E Full-Depth Replacement and Representation Statistics

We compare norms, directions, and accuracy after replacement across all decoder depths in six pretrained model–modality conditions. We then present full TAN trajectories, controls using first-block representations and other audio, results by source, and text-only replacement. Cohorts and aggregation follow Appendix C.

E.1 Norm discrepancy across pretrained models

We compare Qwen3.5-4B and 27B vision, Phi-4 vision and audio, Qwen2-Audio-7B, and Qwen2.5-Omni-3B audio. In all six conditions in Table 7, mean non-text norms exceed text norms at the input and first decoder block. For example, first-block non-text and text norms are 74.45 and 9.49 in Phi-4 vision, and 24.75 and 9.79 in Phi-4 audio. This difference extends to pretrained representations beyond the two models whose formation is tracked in Section 4.1.

Table 7: Original mean L2 norms in pretrained models. Cells give non-text / text, abbreviated NT and T.

Model / input modality	Items	Depth 0 NT / T	Depth 1 NT / T	Depth 3 NT / T
Qwen3.5-4B vision	300	7.10 / 0.72	9.23 / 2.37	8.79 / 3.28
Qwen3.5-27B vision	300	60.41 / 0.92	64.64 / 13.47	102.79 / 17.29
Phi-4 vision	300	74.29 / 11.57	74.45 / 9.49	70.61 / 18.69
Qwen2-Audio-7B	300	34.77 / 0.76	35.31 / 4.28	41.20 / 14.40
Phi-4 audio	256	22.02 / 11.31	24.75 / 9.79	28.05 / 18.71
Qwen2.5-Omni-3B audio	256	34.17 / 0.89	39.88 / 16.71	43.45 / 89.92

In Figure 14, the depths where non-text norms exceed text norms vary by model. For example, Qwen2.5-Omni-3B audio has non-text and text norms of 39.88 and 16.71 in the first block, but 43.45 and 89.92 at depth 3, reversing their relative magnitudes. All models exhibit an input norm difference; subsequent increases in text norms change this relationship. Representation statistics for Phi-4 audio and Qwen2.5-Omni-3B audio use the separate probes described in Appendix C.

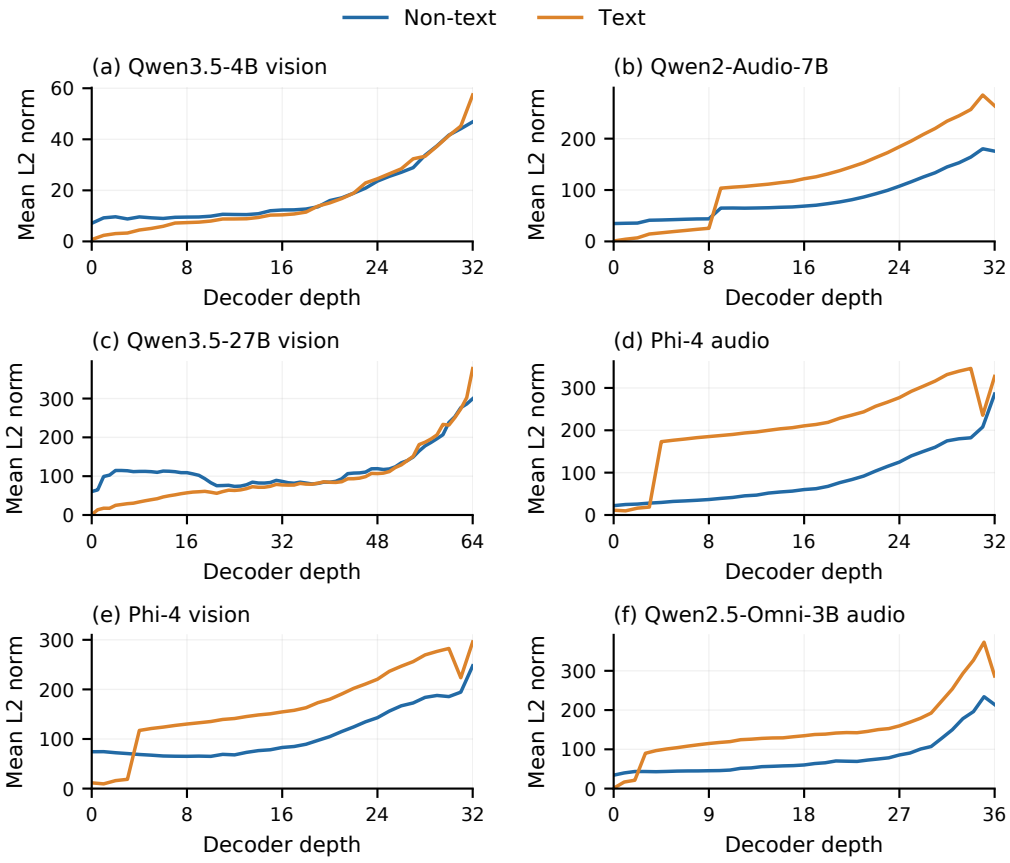


Figure 14: Original mean L2 norms across all depths in six pretrained conditions. Vision is on the left and audio on the right; each panel covers its model's full decoder and norm range.

E.2 Replacement and geometry across decoder depth

Input replacement in six conditions. In Figure 15, Input retains at least 90% of originally correct answers from depth 1 through depth 6 in Qwen3.5-4B vision, depth 3 in Phi-4 vision, and depth 2 in Qwen2-Audio and Qwen3.5-27B vision. Retention declines earlier in the additional audio conditions. Phi-4 audio has 87.7% Input accuracy at depth 1, and Qwen2.5-Omni-3B audio falls from 91.7% at depth 1 to 30.7% at depth 2. In Phi-4 audio, however, depth-1 Norm-matched accuracy is 88.3%, compared with Random's 35.0%. Thus, lower accuracy after replacement with unscaled inputs can coexist with a substantial advantage of input over random directions at the same norm.

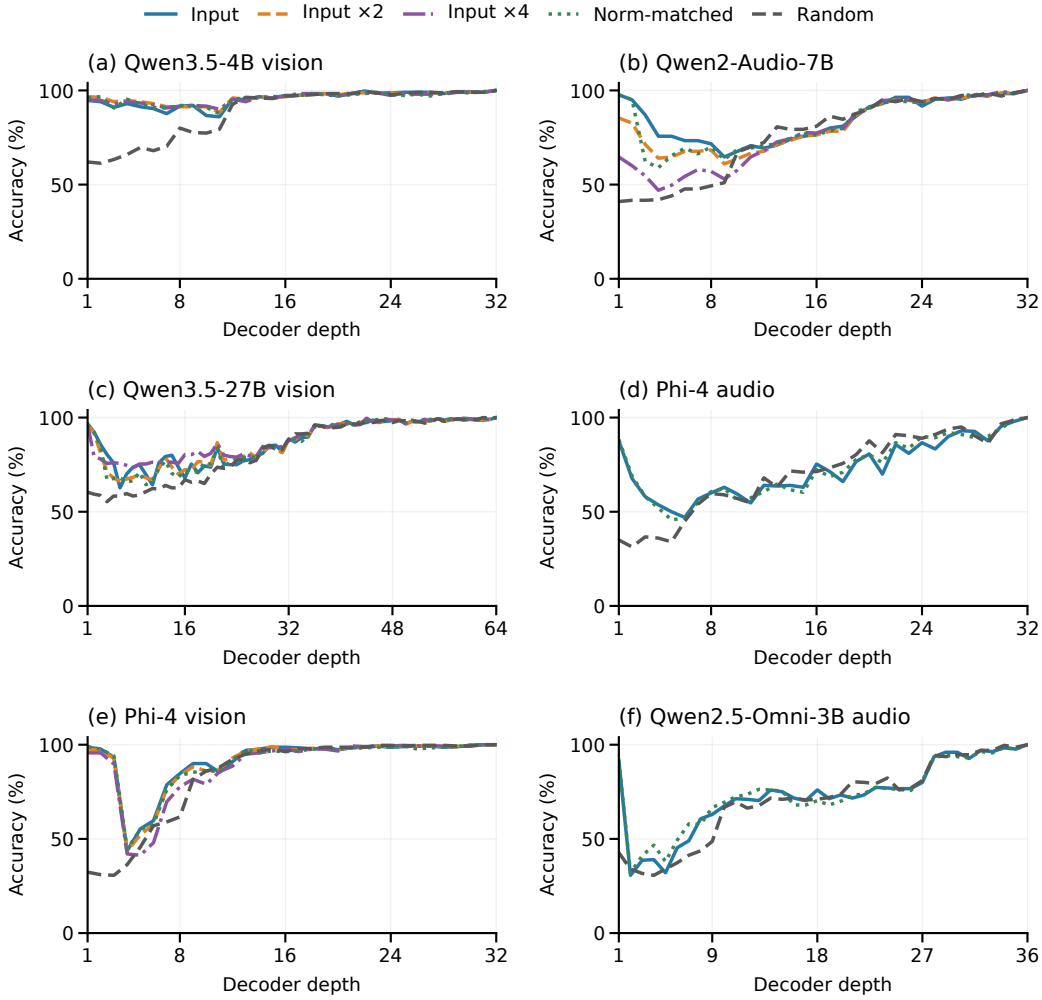


Figure 15: Accuracy after replacement across all depths in six pretrained conditions. Vision is on the left and audio on the right. Phi-4 audio and Qwen2.5-Omni-3B audio measurements cover Input, Norm-matched, and Random.

Directional changes and accuracy at later replacement depths. Input cosine similarity decreases with depth in Figure 16, with the timing and magnitude varying by model. Qwen3.5-27B vision shows a particularly sharp early change. Accuracy after replacement can recover later, alongside high Random accuracy; for example, Qwen3.5-4B vision Random rises from 79.3% at depth 11 to 92.7% at depth 12. At these depths, information transferred to other positions before replacement can sustain predictions across replacement directions. Preservation of the input itself is supported where input cosine similarity remains high and input directions retain an accuracy advantage over Random at the same norm.

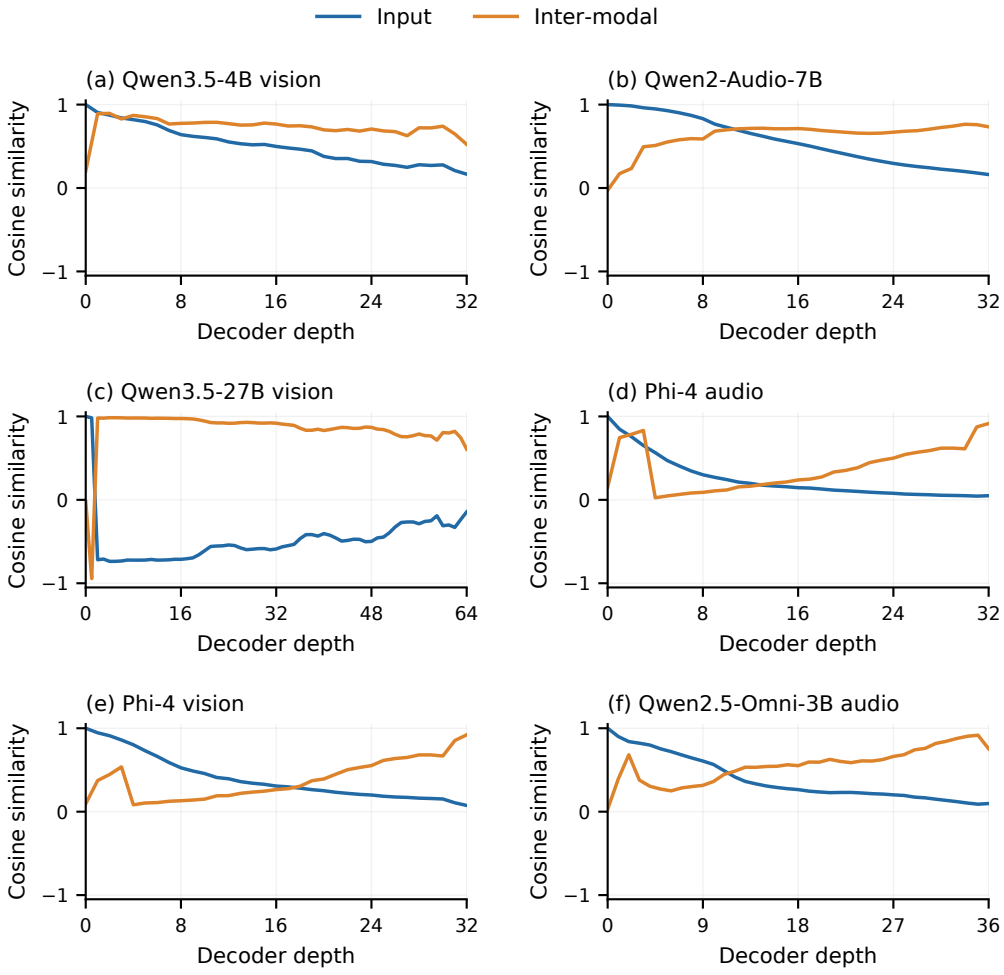


Figure 16: Original cosine similarities across all depths in six pretrained conditions. Input compares tokens with their inputs at the same positions; Inter-modal compares non-text and text mean vectors. Panel order follows Figure 14.

The full decoder after TAN. Figures 17 and 18 compare replacement results and Original representations across all depths for the four training conditions. The replacement curves show early reductions in input preservation and later recovery of accuracy; the Original statistics track norm and direction changes over the same depths. At the final boundary, non-text replacement leaves the already computed response position unchanged, accounting for high accuracy across conditions as described in Appendix C.

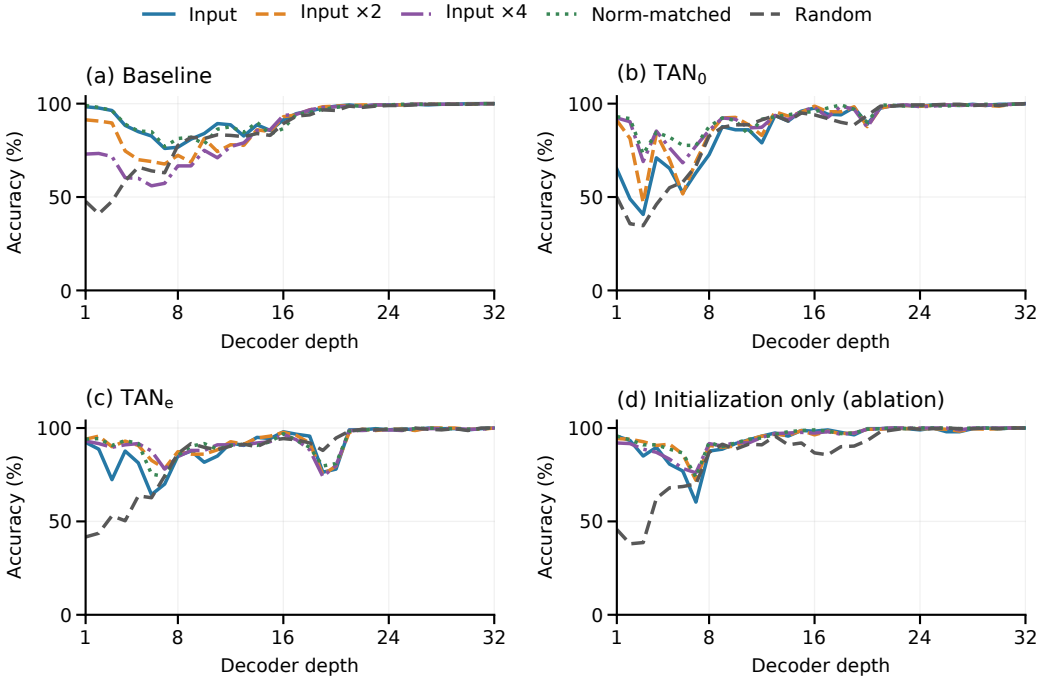


Figure 17: Accuracy after replacement across all depths after Qwen3.5-4B audio alignment.

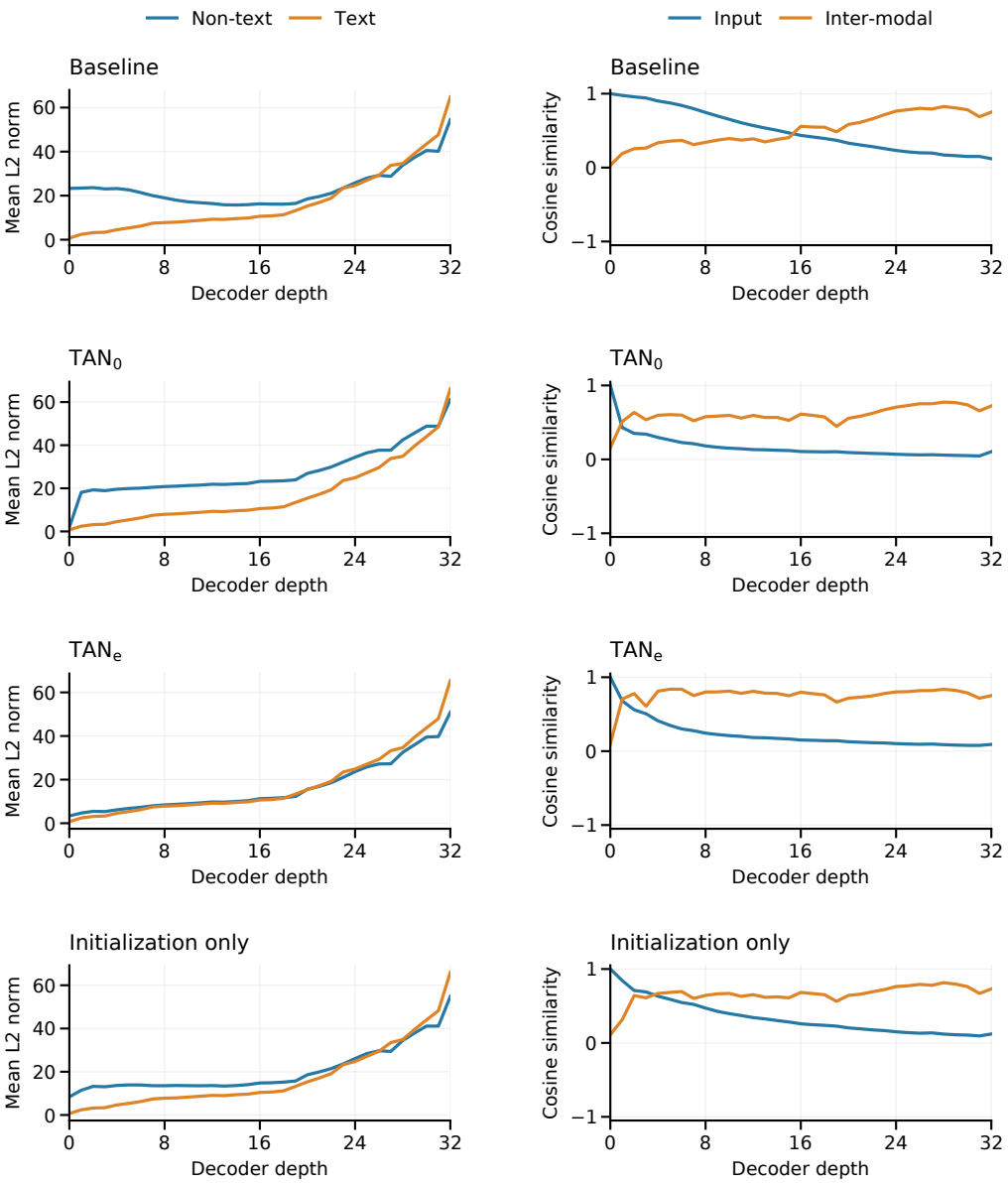


Figure 18: Original representations in the TAN comparison. Left: mean L2 norms. Right: input and inter-modal cosine similarities.

E.3 Norm regularization across training and decoder depth

Norm differences across training steps and depths. Figure 19 extends the input and first-block comparison in Figure 5 across the full decoder. At 10 checkpoints each for short and long training, we measure the difference between mean non-text and text L2 norms on the same 256 probes. The decoder is frozen, and each cell corresponds to one training checkpoint and depth. Both TAN conditions maintain small norm differences at the input, but their internal patterns differ with the extent of regularization. Across all recorded steps at depths 1–11, non-text norms exceed text norms by 4.07–31.39 in TAN_0, whereas differences remain small at -0.19–2.32 in TAN_e. Regularizing norms from the input through the early decoder thus mitigates the formation of large norm differences beyond the first block.

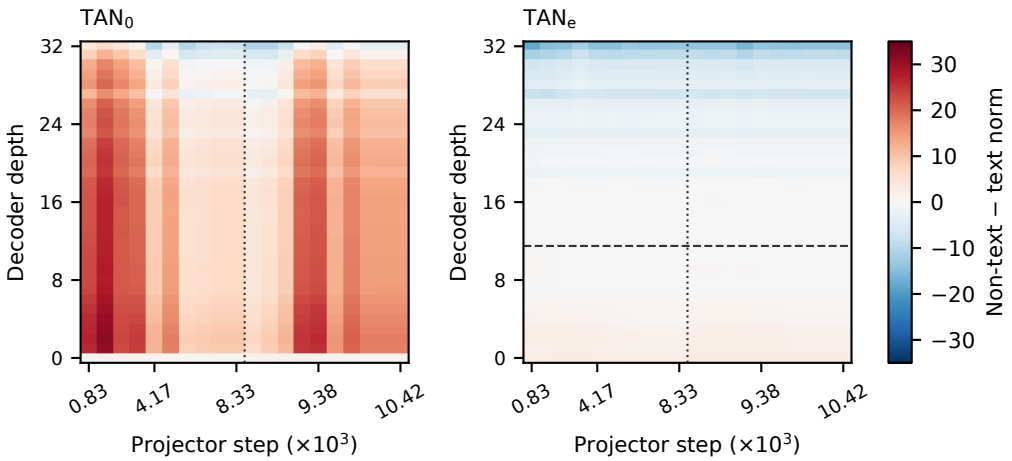


Figure 19: Difference between mean non-text and text L2 norms across all depths and saved checkpoints during TAN projector training. Both panels share a color scale. The vertical dotted line separates short and long training; depths 0–11, below the horizontal dashed line in the right panel, are regularized by TAN_e. Horizontal spacing follows checkpoint order; training-step intervals are unequal.

E.4 REPLACEMENT WITH THE FIRST-BLOCK REPRESENTATION IN TAN_0

Reference representation and controls. We examine how well TAN_0's first-block representation, changed in magnitude and direction, replaces later non-text tokens. Depth-1 outputs serve as references for replacement at the same positions at depths 2–32. Input, Block 1, and Random all receive the Original token norm at the replacement position to compare directions. We use the same 300 commonly correct items, prompts, options, and trained weights as in the main text. The input-direction control matched to the replacement-position norm reproduces the main-text Norm-matched option logits. Random vectors are generated separately for this supplementary experiment.

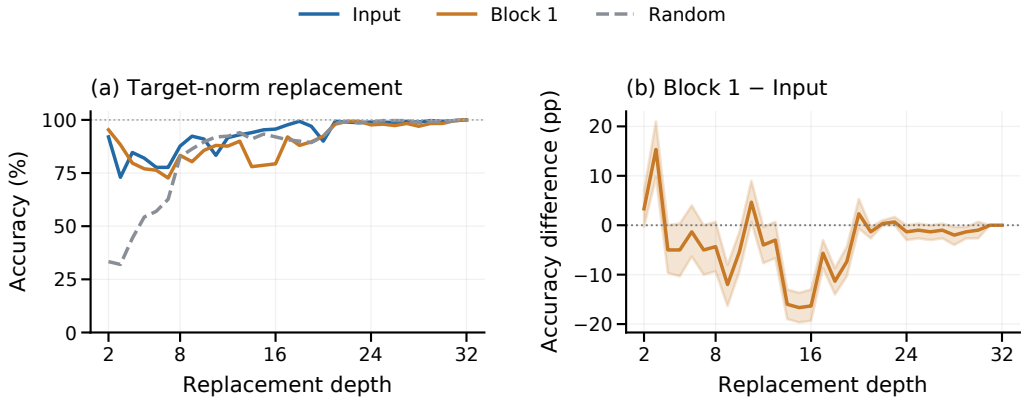


Figure 20: Replacement with TAN_0's first-block representation at all subsequent depths. (a) Accuracy for Input, Block 1, and Random matched to the replacement-position norm, with Original at 100%. (b) Paired Block 1 minus Input accuracy differences and depth-specific 95% confidence intervals from 5,000 cluster bootstrap replicates within each source. The horizontal axis gives the replacement depth; the reference is always depth 1.

Declining accuracy with first-block replacement. In Figure 20, accuracy after replacing tokens with the first-block representation falls from 95.3% at depth 2 to 72.7% at depth 7. Its advantage over Random at the same norm shrinks from 62.0 to 10.0 percentage points. The first-block representation outperforms Input at depths 2–3, but Input is higher at depths 4–7. The difference is 3.3 points $[-0.3, 7.0]$ at depth 2 and 15.3 points [9.7, 21.0] at depth 3. First-block direction contributes to prediction in these early depths, but its ability to replace later tokens decreases with depth.

High Random accuracy at later depths. Random also scores highly over depths 8–20, at 82.7–94.0%. In particular, the first-block representation scores 4.0–14.7 points below Random over depths 9–16, showing regions where this direction impairs prediction when processed with the already computed text states. Over depths 21–32, all three conditions score at least 97%, reaching 100% at the final boundary. Across the full-depth results in Table 8, the first-block representation has a clear advantage over Random at early subsequent depths. Later, information retained outside the replaced positions can sustain high accuracy across replacement directions.

Table 8: Accuracy after replacement (%) and depth-specific 95% confidence intervals for TAN_0 at all subsequent depths. All conditions use the Original norm at the replacement position.

Depth	Input	Block 1	Random
2	92.0 [89.0, 95.0]	95.3 [93.0, 97.3]	33.3 [28.3, 38.3]
3	73.0 [68.0, 77.7]	88.3 [84.7, 91.7]	32.0 [27.3, 36.7]
4	84.7 [80.7, 88.7]	79.7 [76.0, 83.0]	44.3 [38.7, 49.7]
5	82.0 [77.7, 86.3]	77.0 [73.7, 80.3]	54.3 [49.0, 59.7]
6	77.7 [73.0, 82.0]	76.3 [73.0, 79.7]	57.0 [51.7, 62.7]
7	77.7 [73.0, 82.0]	72.7 [69.0, 76.3]	62.7 [57.3, 67.7]
8	87.7 [84.0, 91.0]	83.3 [79.3, 87.0]	82.7 [78.7, 86.7]
9	92.3 [89.3, 95.0]	80.3 [76.7, 84.0]	86.3 [82.7, 90.0]
10	91.0 [87.7, 94.0]	85.7 [82.0, 89.3]	89.7 [86.0, 93.0]
11	83.3 [79.7, 87.0]	88.0 [84.3, 91.3]	92.0 [89.0, 95.0]
12	91.7 [88.7, 94.7]	87.7 [84.7, 90.7]	92.3 [89.3, 95.0]
13	93.0 [90.0, 95.7]	90.0 [87.0, 93.0]	94.0 [91.3, 96.7]
14	94.0 [91.3, 96.3]	78.0 [75.0, 81.0]	91.0 [88.0, 93.7]
15	95.3 [92.7, 97.7]	78.7 [75.7, 81.7]	93.3 [90.7, 96.0]
16	95.7 [93.3, 97.7]	79.3 [76.7, 82.0]	92.0 [89.3, 94.7]
17	97.7 [96.0, 99.3]	92.0 [89.3, 94.7]	90.7 [87.7, 93.3]
18	99.3 [98.3, 100.0]	88.0 [85.0, 90.7]	90.0 [87.0, 92.7]
19	97.0 [95.0, 98.7]	89.7 [86.7, 92.3]	89.3 [86.3, 92.0]
20	90.0 [87.0, 92.7]	92.3 [89.7, 95.0]	92.3 [89.7, 95.0]
21	99.3 [98.3, 100.0]	98.0 [96.3, 99.3]	98.3 [96.7, 99.7]
22	99.0 [97.7, 100.0]	99.3 [98.3, 100.0]	99.3 [98.3, 100.0]
23	98.7 [97.3, 99.7]	99.3 [98.3, 100.0]	98.7 [97.3, 99.7]
24	99.0 [97.7, 100.0]	97.7 [95.7, 99.3]	99.0 [97.7, 100.0]
25	99.0 [97.7, 100.0]	98.0 [96.3, 99.3]	99.7 [99.0, 100.0]
26	98.7 [97.3, 99.7]	97.3 [95.3, 99.0]	99.7 [99.0, 100.0]
27	99.3 [98.3, 100.0]	98.3 [96.7, 99.7]	99.7 [99.0, 100.0]
28	99.0 [97.7, 100.0]	97.0 [95.0, 98.7]	98.7 [97.3, 99.7]
29	99.7 [99.0, 100.0]	98.3 [96.7, 99.7]	99.3 [98.3, 100.0]
30	99.3 [98.3, 100.0]	98.3 [96.7, 99.7]	99.0 [97.7, 100.0]
31	99.7 [99.0, 100.0]	99.7 [99.0, 100.0]	99.7 [99.0, 100.0]
32	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]

E.5 Replacement with other audio inputs

Other audio with matched token counts. The main-text Random control assesses the subsequent function of input directions from the encoder. Here, comparison with other audio processed by the same model examines the contribution of the specific audio's information. For each of the 300 commonly correct main-text items, we sought audio from the same source group with the same audio-token count but a different recording. Among 241 items with candidates, we selected one in a fixed order with seed 122 and evaluated Baseline, Initialization only, and TAN_e. The cohort contains 57 transcript-selection, 57 MMAU, 53 MMAU-Pro, and 74 VoiceBench items; source means receive equal weight. **Other input** inserts the selected audio's depth-0 tokens at the original item's non-text positions, matching each token's norm to the Original norm at the replacement position. We compare it with the item's own Norm-matched input on the same cohort, retaining the original text states. Depth-specific 95% confidence intervals use 5,000 paired bootstrap replicates over 227 clusters linked by original recordings and shared control audio.

Contribution of audio-specific input direction. In Figure 21, the advantage of the item's own input is concentrated in the early decoder. At depth 3, for example, it scores 24–27 points above other inputs in all three training conditions; TAN_e scores 90.8% and 66.8%, respectively. Even with internal norms mitigated, the item's own audio direction sustains higher accuracy at the same norm. At depth 12, all three difference intervals include zero. High accuracy after replacement in this later region can also reflect information passed to other positions before replacement.

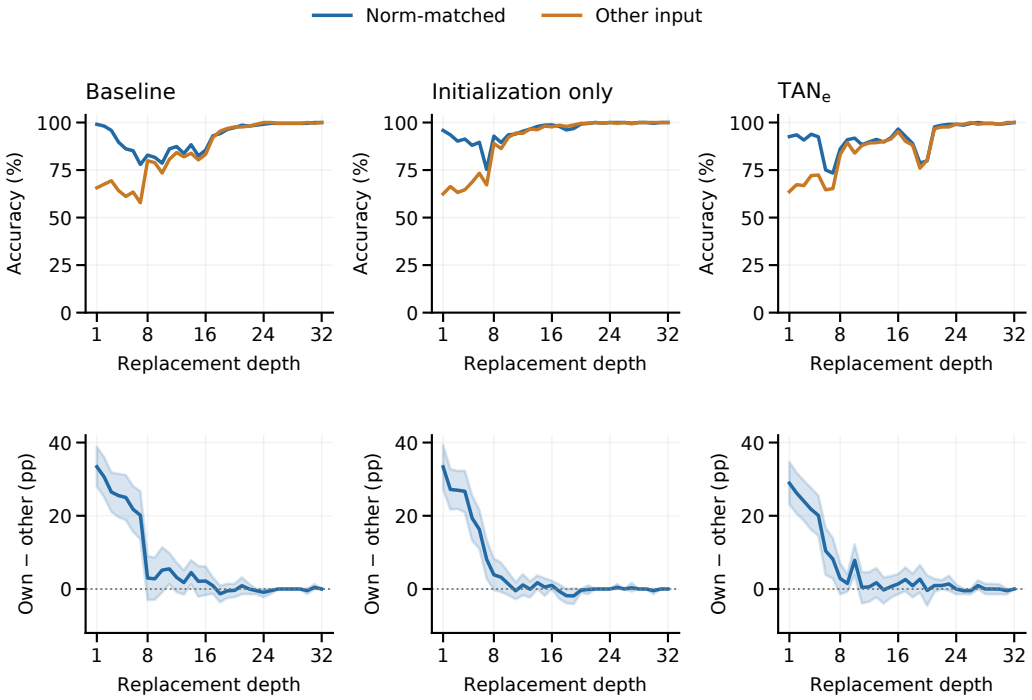


Figure 21: Own-input and other-audio replacement across all depths. Top: accuracy after replacement. Bottom: paired own-input minus other-input differences with depth-specific 95% confidence intervals. Both conditions use the original item's norm at the replacement position. The final depth 32 also checks the computation boundary.

Validation and evaluation scope. We checked weight hashes for the three models and the main-text evaluation items. For each model, own-input Norm-matched option logits at 9,600 item–depth pairs agree with the main-text measurements. The other-input control has zero error in 613 no-op and prefix-replay checks per model, with maximum relative norm error below 0.00196. Other audio inputs pass through the same encoder and trained projector, providing directions closer to the learned input distribution than Random. This supplementary comparison covers items with a token-count-matched candidate; its denominator and source weights differ from the main 300-item accuracy.

E.6 Source-level replacement results

Accuracy changes under norm adjustment also vary by source. Table 9 breaks down the main-text depth-3 results. Each source has 75 items, and their unweighted mean gives overall accuracy. Doubling TAN_e's Input raises transcript-selection accuracy from 46.7% to 97.3% and MMAU from 76.0% to 88.0%; VoiceBench remains at 100.0% in both conditions. Transcript-selection items therefore contribute substantially to the overall recovery at this depth. Figures 15 and 17 show the full-depth trends.

Table 9: Accuracy after replacement (%) by source at depth 3. Matched denotes Norm-matched.

Model / condition	Source	Input	$\times 2$	$\times 4$	Matched	Random
Qwen3.5-4B vision	Procedural	96.0	97.3	97.3	97.3	32.0
Qwen3.5-4B vision	MMBench	97.3	98.7	98.7	96.0	62.7
Qwen3.5-4B vision	MMMU	74.7	82.7	80.0	73.3	69.3
Qwen3.5-4B vision	ScienceQA	96.0	97.3	94.7	96.0	89.3
Baseline	Transcript selection	100.0	93.3	33.3	100.0	32.0
Baseline	MMAU	93.3	89.3	88.0	93.3	68.0
Baseline	MMAU-Pro	92.0	80.0	76.0	92.0	36.0
Baseline	VoiceBench	100.0	96.0	89.3	100.0	54.7
TAN_0	Transcript selection	40.0	38.7	45.3	53.3	32.0
TAN_0	MMAU	72.0	74.7	88.0	82.7	58.7
TAN_0	MMAU-Pro	46.7	56.0	62.7	66.7	37.3
TAN_0	VoiceBench	4.0	18.7	80.0	89.3	10.7
TAN_e	Transcript selection	46.7	97.3	100.0	98.7	34.7
TAN_e	MMAU	76.0	88.0	88.0	86.7	65.3
TAN_e	MMAU-Pro	66.7	74.7	70.7	77.3	37.3
TAN_e	VoiceBench	100.0	100.0	100.0	100.0	74.7
Initialization only	Transcript selection	77.3	100.0	93.3	97.3	33.3
Initialization only	MMAU	92.0	93.3	90.7	93.3	64.0
Initialization only	MMAU-Pro	73.3	78.7	70.7	73.3	36.0
Initialization only	VoiceBench	97.3	98.7	100.0	100.0	21.3

E.7 Text-only replacement with matched information

Input information and replacement positions. To examine whether text inputs also replace later representations, we presented shape attributes or transcripts as text for controlled items from the non-text replacement experiments. The 75 shape items provide counts, colors, and shapes on both sides while retaining the original questions and options. The 75 transcript items provide a recording's transcript and ask for the matching option. Each group is evaluated in three models, yielding 150 unique items and 450 model–item pairs. Shape items have four options in 50 cases and three in 25; all transcript items have three options. Only tokens in the information span are replaced; question and option hidden states and subsequent computation are retained. Input denotes the actual input immediately before the first decoder block. The five conditions in Section 3.2 are applied at every block-output boundary.

Models and measurements. Shape items are evaluated with Qwen3.5-4B, Qwen3.5-27B, and Phi-4; transcript items with Qwen2-Audio-7B, Phi-4, and Qwen2.5-Omni-3B. The two Phi-4 conditions are different tasks on the same backbone; text is processed in LANGUAGE mode with modality LoRA disabled. Qwen3.5-27B uses FP8 and the others BF16. The replaced information spans contain 29 tokens for shapes and 5–55 for transcripts. Text and non-text inputs convey corresponding information through each model's normal input path, so token counts and prompt lengths vary by format. We distinguish accuracy over all 75 items per model and task from retention among Original

correct answers. Table 10 and both figures use accuracy over the same 75 items. The 95% intervals from 5,000 paired bootstrap replicates over item clusters describe sampling uncertainty at each depth.

Table 10: Depth-1 accuracy after replacement (%) for text and non-text inputs conveying the same information. Matched denotes Norm-matched; Original is text accuracy before replacement.

Model / information	Text				Non-text
	Original	Input	Matched	Random	Input
Qwen3.5-4B · shapes	100.0	22.7	21.3	20.0	97.3
Qwen3.5-27B · shapes	100.0	29.3	98.7	32.0	100.0
Phi-4 · shapes	100.0	30.7	30.7	21.3	98.7
Qwen2-Audio · transcripts	97.3	86.7	98.7	37.3	100.0
Phi-4 · transcripts	100.0	34.7	30.7	37.3	100.0
Qwen2.5-Omni-3B · transcripts	100.0	29.3	33.3	37.3	100.0

Effects of adjusting text input norms. In Table 10, replacement with unscaled text inputs describing shapes scores 22.7–30.7%, compared with 97.3–100.0% for non-text inputs on the same items. Matching Qwen3.5-27B text input norms raises accuracy from 29.3% to 98.7%, a paired difference of 69.3 points with a 95% interval of [58.7, 80.0] points. At the same depth, $\times 2$ and $\times 4$ score 28.0% and 38.7%, far below replacement with the target token's norm. For transcripts, Qwen2-Audio likewise rises from 36.0% for Input to 96.0% for Norm-matched at depth 2, with a difference interval of [48.0, 70.7] points. Its Original condition answers 73/75 items correctly; restricted to those items, depth-1 retention is 87.7% for Input and 100.0% for Norm-matched. In contrast, Qwen3.5-4B shape accuracy at depth 1 is low for both Input and Norm-matched, at 22.7% and 21.3%. Input magnitude thus matters for the subsequent function of text representations too, but recovery under norm adjustment varies by model and depth.

High Random accuracy at later depths. Random also achieves high accuracy at later depths in Figures 22 and 23. At depth 16 in Qwen3.5-4B, Input, Norm-matched, and Random all score 100%; information passed to question and option positions before replacement can sustain prediction. At the final boundary, no further information passes between positions, and option logits, NTP losses, and predictions exactly match Original for every replacement condition. The subsequent function of text inputs is therefore evident in early norm-adjustment effects and differences from Random at the same norm. Later recovery is explained by computation that includes information retained outside the replacement positions.

Measurement validation. For every model, reinserting inputs at depth 0 and Original states at all measured depths yields maximum option-logit error at most 10^{-4}. We checked completeness by item, depth, and condition for all 85,950 saved scores and verified predictions from option logits, then recomputed overall accuracy and Original-correct retention from the raw results. Non-text comparisons on the same items were also checked against the corresponding saved scores.

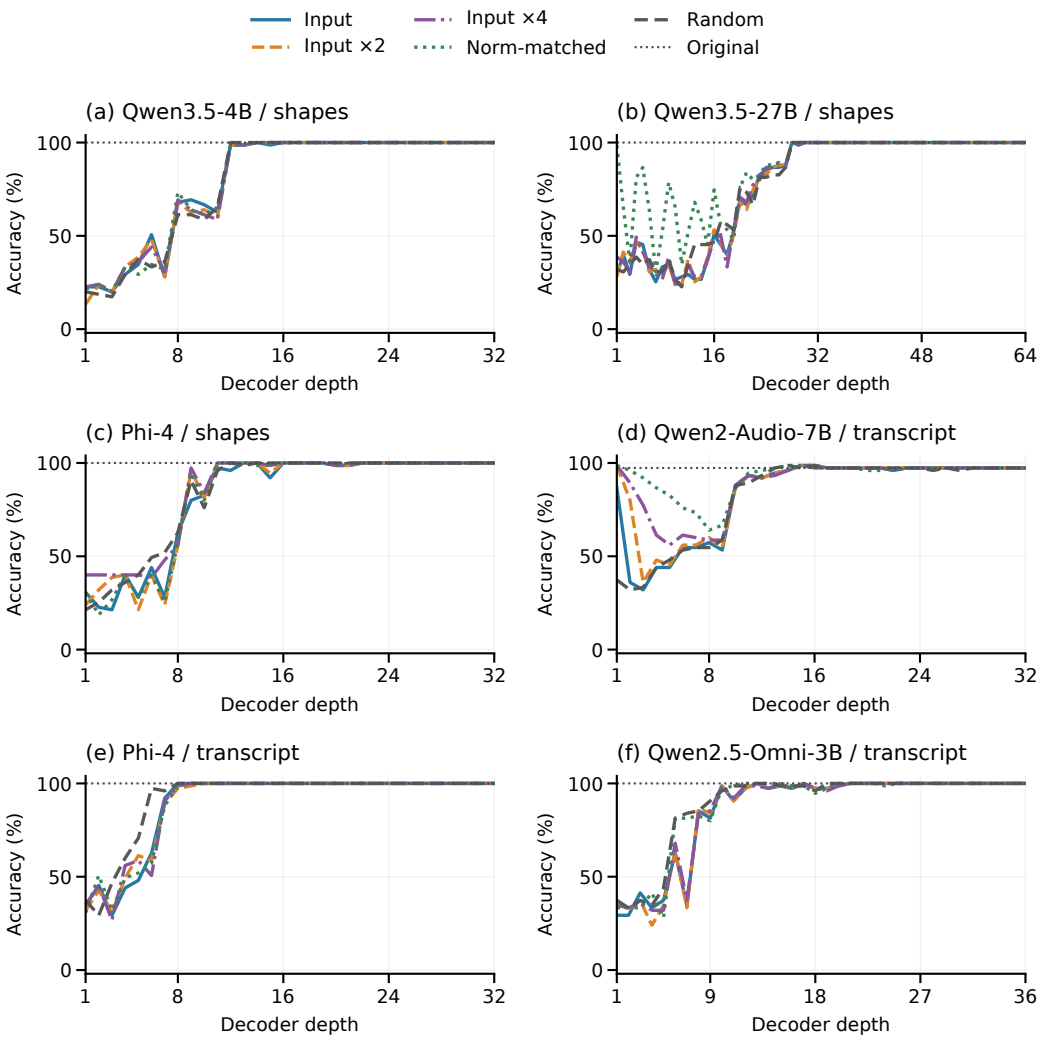


Figure 22: Text-only accuracy after replacement across all depths. Gray dotted horizontal lines indicate Original accuracy.

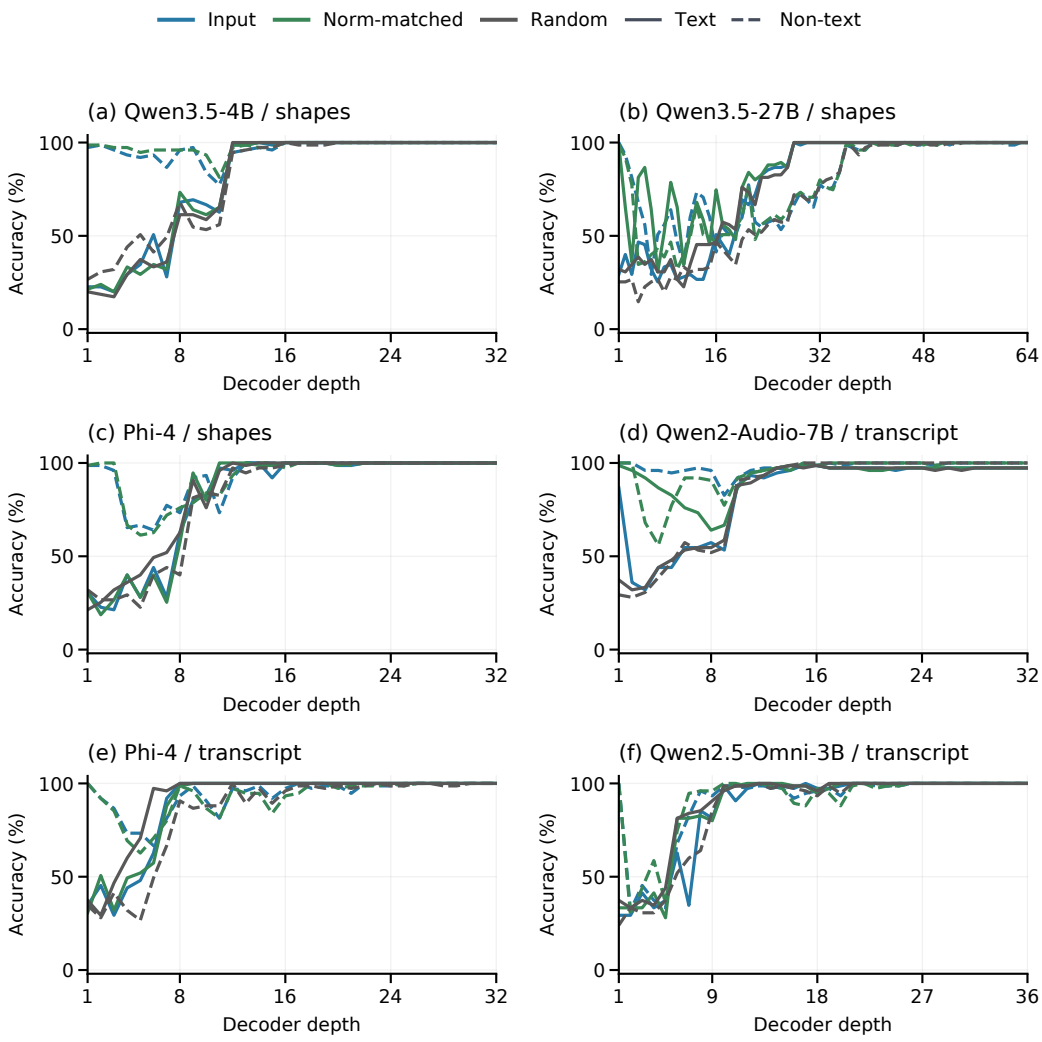


Figure 23: Accuracy after replacement for text and non-text inputs across all depths. Colors indicate replacement conditions; solid and dashed lines indicate input formats.

F BENCHMARK RESULTS AND EVALUATION DETAILS

We compare the four main-text training conditions on transcription, spoken question answering, audio understanding, and classification. Table 11 provides 24 task–metric combinations, including the main-text metrics, additional tasks, and MMAU-Pro category accuracies. We use standard task metrics—accuracy, WER, F1, and exact match—and compare all four conditions on the same evaluation items.

F.1 GENERATION AND SCORING SETTINGS

Generation settings. All four training conditions use BF16 and greedy decoding. Thinking is disabled in the Qwen3.5 chat template, and generation begins after the assistant role header. The maximum number of new tokens is 8,192, subject to the context limit including the input and the model's EOS. The maximum batch size is 32 items, with budgets of 16,000 audio tokens and 65,536 total input and generated tokens. Repetition is checked after at least 128 generated tokens: generation stops if a pattern of at most 512 characters repeats at least eight times and spans at least 512 characters. The response up to termination is passed to the task-specific scorer.

Task instructions. Original questions and supplied options are retained. Where an additional instruction is needed, multiple-choice tasks use "Respond with only the option letter, without explanation.", transcription uses "Respond with only the transcription, without explanation.", and spoken QA uses "Respond with only the answer, without explanation." Existing answer-format instructions in the original prompt take precedence. MMAU-Pro items whose scoring options are not shown to the model use "Respond with only the answer text, without explanation." to generate the answer itself. Instruction-following items retain their original instructions.

MMAU-Pro scorer. Closed-ended responses and their options are embedded with NV-Embed-v2. The option with the highest cosine similarity to the response after L2 normalization is selected as the answer. The scorer's maximum input length is 4,096 tokens. The same model revision and scoring procedure are used for all four training conditions.

F.2 METRIC DEFINITIONS AND DENOMINATORS

WER for LibriSpeech, LibriMix, SpokenSQuAD STT, and Common Voice divides the total word errors by the total reference words. Table n denotes the number of audio examples; WER's denominator is the number of reference words. SpokenSQuAD QA averages per-item exact match and token F1 after response normalization. MMAU and AudioBench (Wang et al., 2025) classification tasks use the accuracy of answers extracted from generated responses. This generation-based scoring differs from option-logit replacement evaluation in Section 3.2.

MMAU-Pro contains 4,559 closed-ended and 87 instruction-following items. Main-text Table 1 divides the combined number of correct answers by 4,646; Table 11 reports the two sets separately and breaks down closed-ended results by category. Thus, TAN_e's main-text accuracy is 35.86%, and its closed-ended accuracy is 36.02%. The same 33 source items with mismatched answers and options are excluded from every condition.

Table scores are the original proportions multiplied by 100. The 95% confidence intervals use the cluster bootstrap in Appendix C. Lower WER and higher accuracy, F1, and EM indicate better performance. Because category sizes differ, overall accuracy is distinct from the unweighted mean of category accuracies.

F.3 TASK SCORES

Table 11 compares the four conditions for each task in one row. Initialization only shares TAN's initialization and learning rate; the two TAN conditions differ in the depths regularized. The table supplements the main tasks with transcription, classification, and MMAU-Pro category results to show task-specific changes.

Table 11: Benchmark scores with 95% confidence intervals. Each cell gives the point estimate and [lower, upper] bounds; n is the number of evaluation items. All scores are multiplied by 100.

Task / metric	n	Baseline	TAN_0	TAN_e	Initialization only (ablation)
Core audio tasks					
LibriMix WER ↓	1,991	9.66 [8.72, 10.84]	8.66 [8.13, 9.17]	**7.70** [7.21, 8.17]	8.11 [7.42, 8.97]
LS test-clean WER ↓	2,620	3.26 [3.06, 3.46]	3.07 [2.85, 3.28]	2.81 [2.61, 3.02]	**2.76** [2.58, 2.95]
LS test-other WER ↓	2,939	7.20 [6.85, 7.54]	7.00 [6.64, 7.35]	**6.26** [5.94, 6.58]	6.35 [6.02, 6.68]
MMAU Acc. ↑	1,000	59.10 [56.20, 62.10]	54.00 [51.00, 57.00]	**69.00** [66.20, 71.80]	59.80 [56.80, 62.80]
SpokenSQuAD QA EM ↑	4,757	67.37 [65.97, 68.78]	75.45 [74.22, 76.75]	**77.55** [76.28, 78.80]	77.02 [75.80, 78.29]
SpokenSQuAD QA F1 ↑	4,757	78.47 [77.35, 79.56]	85.86 [84.99, 86.77]	**87.32** [86.43, 88.18]	86.53 [85.66, 87.41]
SpokenSQuAD STT WER ↓	2,066	15.68 [15.08, 16.34]	**15.51** [14.87, 16.21]	15.96 [15.34, 16.63]	16.32 [15.73, 16.98]
AudioBench					
Common Voice 15 EN WER ↓	16,348	19.63 [18.91, 20.43]	17.99 [17.22, 18.84]	**16.59** [15.89, 17.39]	17.33 [16.27, 18.95]
IEMOCAP emotion Acc. ↑	1,004	23.80 [21.12, 26.39]	**29.88** [27.09, 32.67]	29.38 [26.49, 32.17]	28.78 [26.00, 31.57]
MELD emotion Acc. ↑	2,610	48.81 [46.84, 50.65]	47.78 [45.84, 49.67]	**51.84** [49.90, 53.74]	50.61 [48.66, 52.51]
MELD sentiment Acc. ↑	2,610	**50.77** [48.81, 52.68]	46.32 [44.44, 48.26]	46.21 [44.33, 48.12]	48.20 [46.29, 50.12]
VoxCeleb gender Acc. ↑	4,874	96.96 [96.47, 97.43]	97.54 [97.09, 97.97]	98.42 [98.07, 98.77]	**98.75** [98.42, 99.06]

Table 11 (continued)

Task / metric	n	Baseline	TAN_0	TAN_e	Initialization only (ablation)
MMAU-Pro					
Overall closed-ended Acc. ↑	4,559	34.22 [31.87, 36.25]	34.06 [31.41, 36.26]	**36.02** [33.15, 38.42]	33.87 [31.45, 36.02]
Instruction following Acc. ↑	87	22.99 [14.94, 32.18]	25.29 [16.09, 34.48]	**27.59** [18.39, 36.78]	22.99 [14.94, 32.18]
Multi Acc. ↑	426	**23.00** [18.21, 36.32]	20.42 [14.95, 36.31]	20.19 [14.47, 35.41]	22.30 [16.84, 37.23]
Music Acc. ↑	1,418	36.53 [33.97, 39.02]	40.34 [37.84, 43.00]	**41.33** [38.72, 44.05]	38.58 [36.08, 41.17]
Music + speech Acc. ↑	46	28.26 [16.67, 39.62]	**34.78** [22.00, 48.78]	30.43 [17.31, 46.00]	26.09 [13.95, 39.22]
Sound Acc. ↑	1,027	31.55 [28.66, 34.54]	29.50 [26.81, 32.26]	**33.20** [30.30, 36.13]	29.99 [27.13, 32.97]
Sound + music Acc. ↑	46	**23.91** [11.36, 36.17]	**23.91** [12.77, 36.36]	**23.91** [11.36, 36.73]	17.39 [6.98, 27.66]
Sound + music + speech Acc. ↑	7	**28.57** [0.00, 71.43]	14.29 [0.00, 42.86]	**28.57** [0.00, 57.14]	14.29 [0.00, 42.86]
Sound + speech Acc. ↑	88	31.82 [22.50, 41.00]	**34.09** [22.99, 44.79]	27.27 [18.18, 36.27]	29.55 [19.75, 39.22]
Spatial audio Acc. ↑	321	**38.63** [33.33, 43.81]	37.69 [32.62, 42.90]	33.02 [27.85, 38.10]	33.64 [28.26, 38.82]
Speech Acc. ↑	891	39.62 [36.62, 42.62]	38.38 [35.26, 41.52]	**42.20** [38.90, 45.36]	38.83 [35.75, 42.05]
Voice chat Acc. ↑	289	30.80 [25.61, 36.33]	24.22 [19.38, 29.41]	**33.22** [28.03, 38.41]	32.18 [26.99, 37.37]