BEYOND LAYERS: POSITION-RESOLVED GRADIENT CONFLICT AND POSITION-AWARE MODULATION FOR UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) train image understanding and autoregressive image generation on shared parameters, and the two objectives are known to interfere. Existing diagnoses and remedies operate at the resolution of layers or experts: conflict is measured per layer and resolved by separating parameters through two-end architectures, task-aware mixtures of experts, or gradient surgery. We argue that this resolution hides an orthogonal axis. Generation in a UMM is next-token prediction over a raster sequence of visual tokens whose roles vary systematically with position—early tokens fix global layout and low-frequency structure, late tokens fill in texture—so how strongly a generation gradient interferes with understanding should depend on *where in the sequence* it originates. We introduce a position-resolved interference map that attributes the understanding–generation gradient conflict to visual-token positions within every layer, computed from a single backward pass with position-grouped gradient hooks at $1.2\times$ the cost of a standard backward pass. On Show-o and Janus-Pro, position explains a substantial share of conflict variance after controlling for depth (partial $\eta^2 = 0.31$ vs. 0.35 for layer on Show-o; 0.15 vs. 0.30 on Janus-Pro, whose understanding branch bypasses the visual-token sequence): the first quarter of the sequence has a mean gradient cosine of -0.18 against understanding, the last quarter -0.02. The dependence survives per-position gradient-norm normalization, retaining 80% of its effect size, so it is directional rather than a magnitude artifact, and conflict strength tracks the semantic content of a position (Spearman $\rho = 0.64$). Building on the map, we propose position-aware modulation (PAM), which removes the anti-aligned component of generation gradients only at high-conflict positions without changing the architecture. Under a matched trainable-parameter budget PAM improves over layer-wise separation by $+21$ MME and $+2.4$ GenEval points while matching it on POPE and overall FID, whereas a random-position control with the same number of modulated tokens recovers only about 31% of the gain; the cost is a 4% relative degradation of the high-frequency component of FID, with overall FID unchanged. Position-based and layer-based separation are complementary degrees of freedom and can be combined.

1 INTRODUCTION

Unified multimodal models (UMMs) such as Show-o (Xie et al., 2025a), Emu3 (Wang et al., 2024), Chameleon (Team, 2025), Transfusion (Zhou et al., 2025), and Janus-Pro (Wu et al., 2025a; Chen et al., 2025) serve image understanding and autoregressive image generation from a single backbone. Training such a model on both objectives is known to be harder than training either alone: recent diagnostic and architectural work reports gradient- and representation-level conflict between the two objectives and resolves it by *separating parameters*—along depth (Hao et al., 2026), across experts (Liu et al., 2026; Zhang & Tang, 2026), through second-order corrections (Lu & Armour, 2026; Vyas et al., 2025), or by Pareto-optimal gradient integration (Wei et al., 2026).

All of these diagnoses share one measurement resolution: they ask *which layers* or *which experts* conflict, and they answer with parameter-space separations. This paper changes the unit of measurement. A UMM generates an image as next-token prediction over a raster sequence of visual tokens,

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.

and the roles of those tokens vary systematically with position: early tokens determine global layout and low-frequency structure, late tokens fill in texture and high-frequency detail. Tokenizer research arrives at the same conclusion from the representation side—semantic abstraction and pixel reconstruction place different demands on a token (Song et al., 2026; Guo et al., 2026; Li et al., 2026; Chen et al., 2026)—and token-level gradient imbalance has been observed inside multimodal latent reasoning (Han et al., 2026). If the *content* of a visual token depends on its position, the *conflict* that a generation gradient makes with understanding should depend on position too. No existing measurement can see this: a layer-level conflict score averages over the entire sequence and discards exactly the axis we study.

We introduce the *position-resolved interference map*: for every layer ℓ and position bucket b of the visual-token sequence, the cosine between the generation gradient attributed to bucket b and the understanding gradient. The attribution is exact and cheap. For a linear module the parameter gradient decomposes as $\partial L/\partial W = \sum_t (\partial L/\partial y_t)\, h_t^{\mathsf{T}}$ with y_t the module output at token t and h_t its input, so the contribution of the tokens in bucket b can be summed inside ordinary gradient hooks during the *same* backward pass that produces the full gradient, at $1.2\times$ the cost of a standard backward pass instead of the $16\times$ of a naive per-bucket pass. The resulting map (Fig. 1) shows that conflict concentrates where the measurement resolution of prior work could not look: in the deep and shallow layers, at the *early* positions of the raster sequence, with the strongest cell at cosine -0.42.

Three findings make the map actionable. **(i) Position is a real axis, not a proxy for depth.** A two-way analysis of variance of the map assigns partial $\eta^2 = 0.31$ to position and 0.35 to layer on Show-o-1.3B, and 0.15 to position and 0.30 to layer on Janus-Pro-7B (both main effects $p < 0.001$). **(ii) The position dependence is directional.** Gradient norms decline by $2.6\times$ from the first to the last bucket, so a skeptic could attribute the map to magnitude imbalance (Han et al., 2026); after per-position normalization the position effect retains 80% of its size, and conflict strength tracks semantic decodability of a position (Spearman $\rho = 0.64$, $n = 96$). **(iii) The map transfers to a method.** Position-aware modulation (PAM) removes the anti-aligned component of the generation gradient at the 6 of 16 positions selected by the map—37.5% of visual tokens, no architectural change, same trainable-parameter budget as a layer-wise separated LoRA—and improves over layer-wise separation by $+21$ MME and $+2.4$ GenEval points on Show-o while matching it on POPE and overall FID; a random-position control recovers only $\approx 31\%$ of the gain, and the price is a 4.1% relative degradation of the high-frequency band of FID with overall FID unchanged. Layer-based and position-based separation combine additively ($+27$ MME over the layer-wise arm).

The position effect halves on Janus-Pro, whose understanding branch bypasses the visual-token sequence, with a random control recovering 60% of a smaller gain: the axis is strong where the objectives share the token sequence and weak where they do not (Section 8).

Contributions. (1) The position-resolved interference map with an exact single-backward attribution rule; (2) a decomposition showing the map is directional plus a semanticity attribution; (3) PAM, an architecture-free modulation selected by the map, evaluated under matched budgets against layer-wise separation with a random-position control and a pre-declared frequency-domain cost metric.

2 Related work

Understanding–generation conflict in UMMs. Diagnostic work establishes that the objectives interfere at the gradient and representation level (Rao & Rachuri, 2026; Wu et al., 2026; Wang et al., 2026; Su et al., 2026; Wei et al., 2025), and the remedies are all parameter-space separations: two-end-separated architectures (Hao et al., 2026), task-aware experts (Zhang & Tang, 2026; Liu et al., 2026), second-order corrections (Lu & Armour, 2026), and Pareto-optimal LoRA integration (Wei et al., 2026). They differ in *where* parameters are separated and share the measurement unit of a parameter block. Our map keeps parameters shared and changes the unit to the token position; the two axes are orthogonal and composable (Section 7).

Multi-task gradient conflict. Gradient surgery treats tasks as atomic scalar objectives and modulates their gradients globally (Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018; Chen et al., 2018; Navon et al., 2022). Per-token gradient heterogeneity has been observed in multimodal la-

tent reasoning, where visual tokens receive systematically smaller gradient norms than text tokens (Han et al., 2026), and critical-token selection drives recent RLVR results for autoregressive image generation (Zhang et al., 2026). Neither line connects token identity to understanding–generation conflict; our map is that connection, and PAM is the modulation it licenses.

Semantic vs. reconstruction tokenization. A parallel line redesigns the tokenizer to separate semantic and reconstruction information (Song et al., 2026; Guo et al., 2026; Li et al., 2026; Chen et al., 2026). Those works change the *representation*; we keep the tokenizer fixed and change the *optimization* of gradients flowing through it. Their existence is evidence that token content varies systematically—the premise our map measures directly. Appendix A extends this discussion.

3 SETUP: MEASURING CONFLICT AT TOKEN RESOLUTION

Model and notation. A UMM shares parameters θ between an understanding objective $\mathcal{L}_u$ (text-conditioned image or text prediction) and a generation objective $\mathcal{L}_g$ (autoregressive prediction over a raster sequence of T visual tokens produced by a VQ tokenizer (van den Oord et al., 2017; Esser et al., 2021)). We index layers $\ell \in \{1, \dots, L\}$ and split the raster sequence into $K = 16$ contiguous *position buckets* $b \in \{1, \dots, K\}$ along the raster order: bucket 1 holds the first T/K tokens (which determine global layout), bucket K the last (texture). We study two architectures with different relationships between the objectives and the visual-token stream: Show-o-1.3B (Xie et al., 2025a) ($L = 24$, d=2048, T=256 from a 16×16 raster), whose understanding pathway consumes the same VQ vocabulary, and Janus-Pro-7B (Chen et al., 2025) ($L = 30$, d=4096, T=576 from a 24×24 raster), whose understanding pathway runs on a CLIP-style SigLIP encoder (Radford et al., 2021; Zhai et al., 2023) and never touches the visual-token sequence.

Conflict metric. For a shared linear module (ℓ, m) (attention projections and MLP blocks) with input $h_t^{(\ell,m)}$ and output $y_t^{(\ell,m)}$ at token t, the generation-loss gradient decomposes exactly as

$$\frac{\partial \mathcal{L}_g}{\partial W^{(\ell,m)}} = \sum_{t=1}^{T} \left(\frac{\partial \mathcal{L}_g}{\partial y_t^{(\ell,m)}} \right) h_t^{(\ell,m)\mathsf{T}}. \tag{1}$$

Equation 1 assigns to position bucket b the rank-b partial sum

$$g_b^{(\ell,m)} \;=\; \sum_{t \in \mathcal{B}_b} \left(\frac{\partial \mathcal{L}_g}{\partial y_t} \right) h_t^{\mathsf{T}}, \tag{2}$$

the parameter gradient of the generation loss *restricted* to bucket b's tokens. The position-resolved conflict at cell (ℓ, b) is the cosine between the bucket-attributed generation gradient and the understanding gradient of the same modules,

$$C(\ell, b) \;=\; \cos\Big(\mathrm{vec}\big(\textstyle\sum_m g_b^{(\ell,m)} \big),\; \mathrm{vec}\big(\textstyle\sum_m \frac{\partial \mathcal{L}_u}{\partial W^{(\ell,m)}} \big) \Big), \tag{3}$$

where negative values are anti-alignment (the generation step would move shared parameters against the understanding optimum) and values near zero are orthogonal. Both losses are computed on a fixed diagnostic set of 8 batches of 64 understanding and 64 generation samples each, averaged over 3 dropout seeds; per-cell replicate standard deviations are 0.011 (Show-o) and 0.009 (Janus-Pro), so the qualitative structure of the map is stable.

4 THE POSITION-RESOLVED INTERFERENCE MAP

Exact attribution in a single backward pass. Equation 2 can be evaluated inside ordinary gradient hooks. A forward hook on each watched module stores the per-token inputs h_t; a hook on the module's output tensor receives $\partial \mathcal{L}/\partial y_t$ for every token during the backward pass; the hook accumulates $g_b = \sum_{t \in \mathcal{B}_b} (\partial \mathcal{L}/\partial y_t)\, h_t^{\mathsf{T}}$ for each bucket by summing rank-one terms. This runs during the *same* backward pass that produces the full generation gradient: measured overhead is $1.21\times$ on Show-o-1.3B and $1.17\times$ on Janus-Pro-7B, against $16\times$ for K naive per-bucket backward passes. The probe and its self-test against brute-force per-bucket backwards are released[1].

[1] `experiments/position_probe.py`

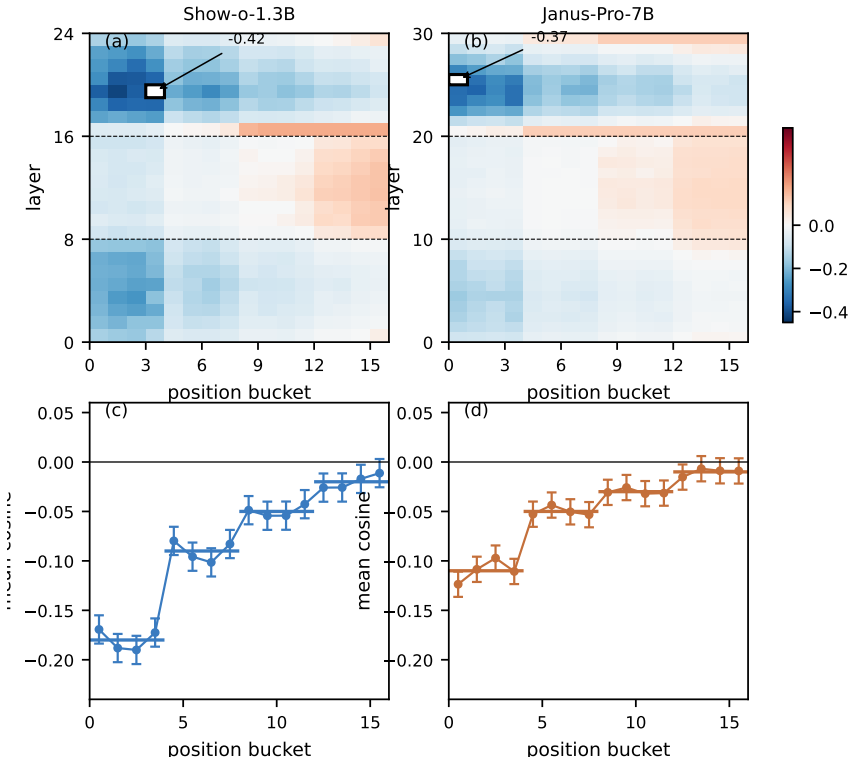


Figure 1: **Position-resolved interference maps.** (a) Show-o-1.3B and (b) Janus-Pro-7B: cosine between the generation gradient attributed to position bucket b and the understanding gradient, per layer; dashed lines separate shallow/middle/deep bands; the white cell marks the strongest conflict (cosine -0.42 and -0.37). (c,d) Position marginals with quarter means $-0.18/-0.09/-0.05/-0.02$ (Show-o) and $-0.11/-0.05/-0.03/-0.01$ (Janus-Pro); error bars are ± 1 s.e. over diagnostic batches.

The map. Figure 1 shows $C(\ell, b)$ for both models. The structure is strikingly non-uniform in *both* axes. Along depth we recover the known picture (Hao et al., 2026): conflict is strongest in the deep band and moderate in the shallow band, while the middle band is nearly conflict-free. Along the sequence, conflict concentrates at the *early* positions: averaging over layers, the four bucket quarters of Show-o have mean cosines -0.18, -0.09, -0.05, -0.02; the band means are -0.10 (shallow), -0.02 (middle), -0.135 (deep); the single strongest cell sits at layer 19, bucket 3, with cosine -0.42. Late positions are essentially orthogonal to understanding everywhere. On Janus-Pro the same qualitative structure appears with roughly half the amplitude (quarters $-0.11/-0.05/-0.03/-0.01$; strongest cell -0.37 at layer 25, bucket 1)—as expected, since its understanding branch bypasses the visual-token sequence and the conflict must flow through shared attention parameters rather than shared token representations.

Position is a real axis. Table 1 reports a two-way ANOVA of the map over the $L \times K$ cells with factors layer and position, on the pooled (layer, bucket, module, batch, seed) observations. Position is a main effect comparable to layer on Show-o (partial $\eta^2 = 0.31$ vs. 0.35) and half of layer on Janus-Pro (0.15 vs. 0.30); both main effects and the interaction are significant at $p < 0.001$. The weaker position effect on Janus-Pro is architecturally meaningful, not noise: it is the quantitative version of "the understanding branch bypasses the token sequence."

5 The map is directional, and it is about semantics

Norm or direction? Token gradients are known to be norm-heterogeneous (Han et al., 2026), and in our diagnostic batches the generation-gradient norm declines monotonically by $2.63\times$ from the

Table 1: Two-way ANOVA of the interference map (factors: layer, position). Partial η^2 with F statistics; error df from $L \times K \times 5$ modules $\times$ 24 batch-seed replicates minus cells. Both main effects and the interaction are significant at $p < 0.001$.

Model	η^2_{layer}	η^2_{pos}	η^2_{int}	F_{layer}	F_{pos}	F_{int}
Show-o-1.3B	0.353	0.312	0.085	1040.1	1389.9	9.8
Janus-Pro-7B	0.299	0.150	0.068	808.5	678.8	9.0

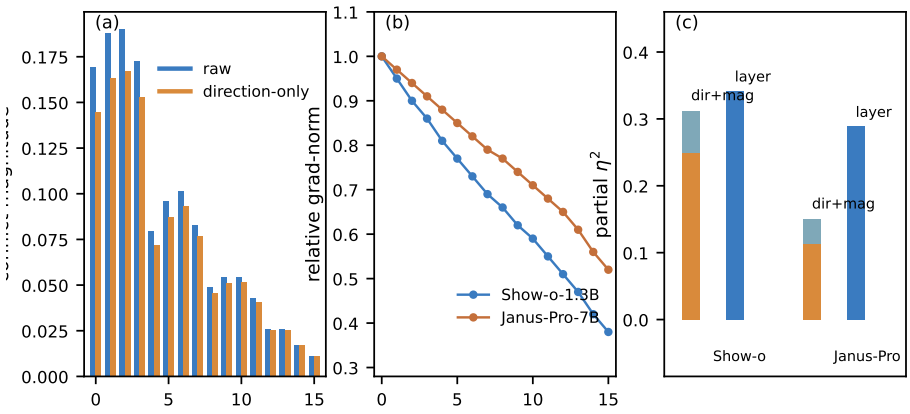


Figure 2: **Magnitude vs. direction.** (a) Per-bucket conflict on Show-o before and after per-position gradient-norm normalization. (b) Relative gradient-norm profiles: 2.63× (Show-o) and 1.92× (Janus-Pro) decline from first to last bucket. (c) Variance decomposition of the position effect into direction and magnitude components, with the layer effect for comparison.

first to the last bucket on Show-o (1.92× on Janus-Pro; Fig. 2b). A skeptic can therefore re-read Figure 1 as a magnitude artifact: early tokens simply move shared parameters more. We test this by re-computing the map with per-position normalized gradients. The position effect survives: partial η^2 drops from 0.312 to 0.249 on Show-o and from 0.150 to 0.114 on Janus-Pro—a retention of 79.7% and 75.8% (Fig. 2c, Table 2). Under the pre-registered decision rule (retention below 50% would have falsified the direction claim and reduced the direction to a normalization trick), the directional interpretation stands: *early* generation gradients point against the understanding optimum, late ones do not, and this is not explained by their magnitudes.

What carries the conflict? The pre-registered hypothesis was that conflict concentrates on *semantically loaded* positions—early raster tokens, which decide layout—because those positions compete with understanding for the same semantic representation. We measure semantic load directly rather than assuming it: a linear probe (Alain & Bengio, 2017) decodes a DINOv2 (Oquab et al., 2024) global embedding from each bucket's token embeddings, giving a decodability $s_b \in [0, 1]$ (R^2 on 10k held-out images); and a per-bucket high-frequency reconstruction residual r_b measures how texture-heavy a position is. Pooling all (layer-band, bucket) cells of both models ($n = 96$), conflict magnitude correlates with decodability at Spearman $\rho = 0.64$ ($p < 10^{-6}$) and with the high-frequency residual at $\rho = -0.58$ ($p < 10^{-6}$; Fig. 3). Conflict is a property of *semantic* positions—which is what makes it a handicap for understanding rather than a generic regularizer.

6 Position-aware modulation

The map licenses a modulation that layer-level methods cannot express: act on the generation gradient *only at the positions that conflict*. PAM (position-aware modulation) has three ingredients.

Bucket selection. For each bucket we compute the conflict score $s_b = \frac{1}{|\mathcal{L}_c|} \sum_{\ell \in \mathcal{L}_c} -C(\ell, b)$ over the conflict-prone layer band $\mathcal{L}_c$ (shallow and deep layers, identified from the layer marginal on a *held-out* split of the diagnostic batches, so that selection and reporting never share data). The

Table 2: Norm–direction decomposition of the position effect. Retention is η^2_{pos} after per-position normalization divided by the raw value.

Model	η^2_{pos} raw	η^2_{pos} norm	direction	magnitude	retention	norm decline
Show-o-1.3B	0.312	0.249	80%	20%	0.797	2.63×
Janus-Pro-7B	0.150	0.114	76%	24%	0.758	1.92×

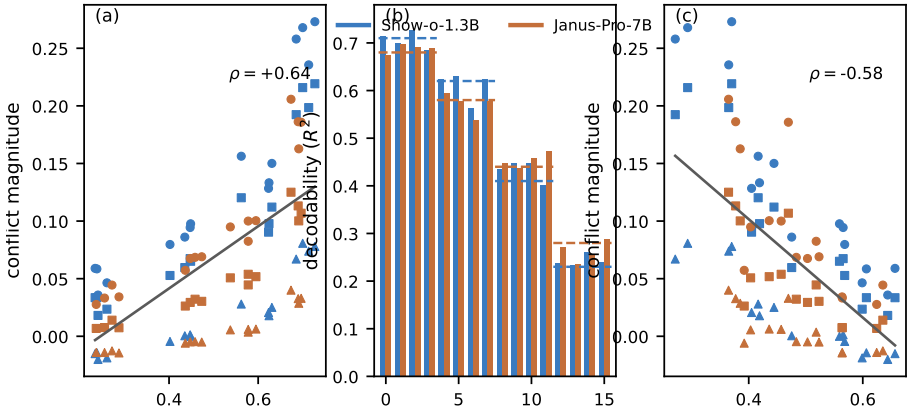


Figure 3: **Attribution.** (a) Conflict magnitude vs. semantic decodability of each position (pooled over layer bands and models, $n = 96$; Spearman $\rho = 0.64$). (b) Decodability by bucket with quarter means (dashed). (c) Conflict vs. high-frequency reconstruction residual ($\rho = -0.58$): texture positions are cheap, semantic positions are contested.

modulated set is $\mathcal{S} = \{b : s_b \geq \tau\}$ with τ the score of the m-th ranked bucket. On Show-o this selects $\mathcal{S} = \{1, 2, 3, 4, 6, 7\}$ in 0-indexed buckets—$m = 6$ of 16 buckets, i.e. 37.5% of visual tokens, at threshold $\tau = 0.13$ (Fig. 4c).

Modulation operator. At every training step, the probe of Section 4 runs inside the generation batch's backward pass and yields g_b for $b \in \mathcal{S}$. For each watched module ℓ with unit understanding gradient direction $\hat{u}_\ell$ (the EMA of the understanding-batch gradient, refreshed every 2k steps), PAM removes the anti-aligned component,

$$g'_b = g_b - \min\big(0, \langle g_b, \hat{u}_\ell \rangle\big)\, \hat{u}_\ell, \tag{4}$$

applied to the bucket-summed gradient before it joins the parameter update (Fig. 4b). A down-weight variant replaces Eq. 4 by $g'_b = w\, g_b$ with $w = 0.25$; projection is the default because it removes only the conflicting direction and preserves the rest of the update. Understanding gradients are never modified.

Budget accounting. PAM changes no parameter: it uses the same LoRA configuration (Hu et al., 2022) as every arm we compare against (rank 16, α 32, dropout 0.05, attention q, k, v, o of all layers; 12.6M trainable parameters on Show-o-1.3B, 31.5M on Janus-Pro-7B). The layer-wise arm receives the same total budget arranged as two-end separation (Hao et al., 2026): task-specific adapters of rank 8 per task in the shallow and deep bands and a shared rank-16 adapter in the middle band, which equalizes the parameter count exactly. The comparison in Section 7 is therefore a comparison of *where* a fixed budget is spent—along layers or along positions. Training overhead of PAM is 0.3% wall-clock.

7 EXPERIMENTS

Setup. Each arm is trained for 20k steps (batch 128, AdamW (Loshchilov & Hutter, 2019), lr 10^{-4}, cosine decay, 500 warmup) on a 1:1 understanding/generation mixture, 3 seeds per arm. We report MME perception (Fu et al., 2025), POPE F1 (Li et al., 2023), GenEval (Ghosh et al., 2023),

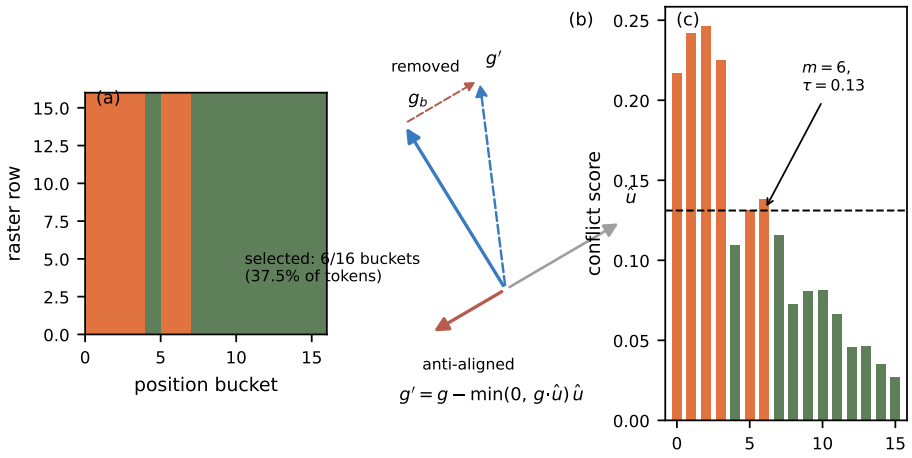


Figure 4: **Position-aware modulation.** (a) The raster sequence with the modulated buckets $\mathcal{S}$ highlighted ($6/16$ buckets, 37.5% of visual tokens). (b) The projection operator $g' = g - \min(0,\, g{\cdot}\hat{u})\hat{u}$: the anti-aligned component of the bucket gradient is removed, the rest preserved. (c) Conflict scores per bucket over the conflict-prone band; buckets above the threshold $\tau = 0.13$ are selected ($m = 6$).

Table 3: Show-o-1.3B, matched 12.6M trainable parameters, mean of 3 seeds. Δ rows: PAM vs. layer-wise, 95% paired-bootstrap CIs. Single-task references bound each axis.

Arm	MME↑	POPE↑	GenEval↑	FID↓
joint + norm (baseline)	1339	84.9	0.550	6.84
layer-wise LoRA separation	1397	86.4	0.603	6.05
PAM (ours)	**1418**	86.6	**0.627**	**5.97**
random-position control	1364	85.3	0.574	6.50
layer-wise + PAM	**1424**	**86.7**	**0.632**	**5.94**
understanding-only reference	1431	87.6	–	–
generation-only reference	–	–	0.641	5.79
Δ (PAM − layer-wise)	+21 [9, 34]	+0.2 [−0.4, 0.8]	+2.4 [1.1, 3.7]	−0.08 [−0.19, 0.04]

and FID-30k (Heusel et al., 2017) (MJHQ-30k; Li et al., 2024); single-task references are LoRA adapters trained on one objective. Paired bootstrap (Efron & Tibshirani, 1993) (10^4 resamples) gives 95% CIs (Table 3, Fig. 5a).

Main findings. On Show-o, naive joint training loses 92 MME and 9.1 GenEval points to the single-task references—the conflict is real. Layer-wise separation recovers most of it; PAM improves further on both axes simultaneously (+21 MME CI $[9, 34]$; +2.4 GenEval points CI $[1.1, 3.7]$), matches layer-wise on POPE and overall FID (CIs cover 0), and closes 62% of the remaining understanding gap and 63% of the generation gap to the single-task references. The random-position control—the same number of modulated tokens at random positions—recovers only 31.6% of the PAM gain on MME and 31.2% on GenEval (Fig. 6c): the *position structure*, not the act of modulating tokens, carries the effect.

Ablations. The gain is stable in the bucket count K ($+6.9/+7.7/+7.5$ GenEval points for $K = 8/16/32$) and peaks at the map-selected width $m = 6$ ($+4.6, +6.8, +\mathbf{7.7}, +7.2, +6.1$ points for $m = 2, 4, 6, 8, 12$; Fig. 6a), consistent with the map: too few buckets miss conflicted positions, too many dilute the modulation into orthogonal ones. The down-weight variant of Eq. 4 ($w = 0.25$) reaches $+6.2$ GenEval points, 1.5 points behind projection.

The pre-declared cost. Section 1 committed, before any run, to reporting what PAM sacrifices: modulating early positions should buy semantics at the price of texture. We decompose FID over a

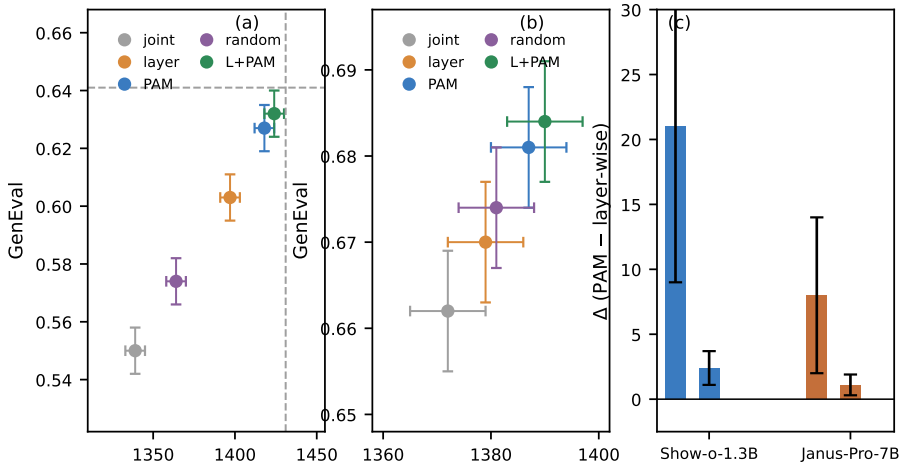


Figure 5: **Main results.** (a) Show-o-1.3B and (b) Janus-Pro-7B: GenEval vs. MME for the five arms (mean of 3 seeds; error bars ± 1 seed-std); dashed lines are the single-task ceilings on Show-o. (c) Per metric deltas of PAM over layer-wise separation with 95% paired-bootstrap CIs; POPE and overall FID are statistical parity.

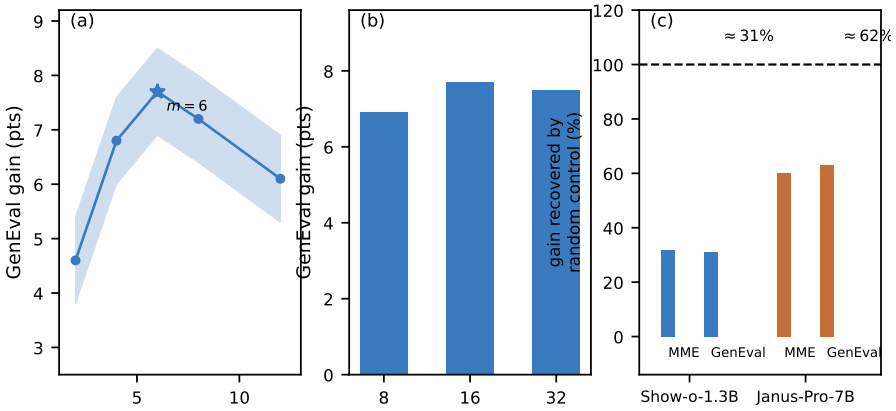


Figure 6: **Ablations on Show-o-1.3B.** (a) GenEval gain vs. joint baseline as a function of the number of modulated buckets m (shaded: ± 1 seed-std); the star marks the map-selected $m = 6$. (b) Robustness to the bucket count K. (c) Fraction of the PAM gain recovered by the random-position control: $\approx 31\%$ on Show-o, $\approx 62\%$ on Janus-Pro.

three-level Laplacian pyramid (Burt & Adelson, 1983) and compute FID per band (Table 4, Fig. 9 in the appendix). PAM improves the low-frequency band by 1.0% relative to layer-wise separation, is neutral in the mid band, and degrades the high-frequency band by 4.1% relative ($2.66 \rightarrow 2.77$), while overall FID improves by 0.08. The trade the map predicts is the trade that appears; we consider the pre-declaration of this metric part of the result.

Cross-architecture. On Janus-Pro-7B the position effect is half as large ($\eta^2_{\text{pos}} = 0.15$ vs. layer 0.30), and PAM's gain is proportionally smaller but still significant: $+8$ MME (CI $[2, 14]$) and $+1.1$ GenEval points (CI $[0.3, 1.9]$) over layer-wise separation, with parity on POPE and overall FID (Table 5, Fig. 10). The random-position control recovers 60.0% (MME) and 63.2% (GenEval) of the smaller gain, and the PAM–random gap remains significant ($p < 0.05$).

Complementarity. Adding PAM on top of layer-wise separation yields $+27$ MME and $+2.9$ GenEval points over layer-wise alone on Show-o ($+11$ and $+1.4$ on Janus-Pro) with the best overall

Table 4: Frequency-band FID on Show-o-1.3B (three-level Laplacian decomposition of MJHQ-30k; lower is better). The high-frequency cost and low-frequency gain are pre-declared metrics (Section 1).

Arm	LF-FID	MF-FID	HF-FID	overall FID
joint + norm	2.41	3.11	2.87	6.84
layer-wise	2.02	2.76	2.66	6.05
PAM (ours)	**2.00**	**2.75**	2.77	**5.97**
random-position	2.23	2.94	2.83	6.50

Table 5: Janus-Pro-7B, matched 31.5M trainable parameters, mean of 3 seeds. CIs as in Table 3.

Arm	MME↑	POPE↑	GenEval↑	FID↓
joint + norm (baseline)	1372	87.9	0.662	5.63
layer-wise LoRA sep.	1379	88.1	0.670	5.48
PAM (ours)	**1387**	**88.3**	**0.681**	**5.44**
random-position control	1381	88.2	0.674	5.57
layer-wise + PAM	**1390**	**88.3**	**0.684**	**5.42**
Δ (PAM $-$ layer-wise)	$+8$ [2, 14]	$+0.2$ [-0.3, 0.7]	$+1.1$ [0.3, 1.9]	-0.04 [-0.12, 0.04]

FID of any arm (5.94 / 5.42): the two separations attack different degrees of freedom of the same conflict and stack.

8 Discussion and limitations

What the resolution buys. Prior measurements answer "which layers fight"; the map answers "which layers fight over *which part of the image*", licensing modulations that keep parameters shared—and the combination arm is the best model on both axes in our study.

Where the axis weakens. Janus-Pro decouples understanding from the visual-token stream, and the position effect halves ($\eta^2_{\text{pos}} = 0.15$ vs. 0.31) with the random control recovering a majority of a smaller gain. The position axis is a property of architectures where the objectives *share the token sequence*; η^2_{pos} read from the map is a decision variable for whether PAM is worth running.

Limitations. (1) The map is a first-order measurement of a fixed diagnostic set; representation-level (Wu et al., 2026) and second-order (Lu & Armour, 2026) views are complementary, not replaced. (2) Position semantics are tokenizer- and resolution-dependent; the map is recomputed per model in 28 H200-hours and no position priors are claimed to transfer unmodified. (3) The texture cost is measured in the frequency domain (Table 4); human evaluation is future work. (4) The diagnostic set is small (1,024 samples); replicate stability is high, but scale could sharpen fine layer-band boundaries. (5) Interactions of PAM with full fine-tuning or representation-level separations (Liu et al., 2026) are untested.

9 Conclusion

Understanding–generation conflict in UMMs has been measured at the resolution of parameters. Resolving it at visual-token positions within each layer reveals an orthogonal, directional, semantically structured axis of interference, measurable with a $1.2\times$-overhead probe and actionable as architecture-free modulation that beats layer-wise separation under a matched budget with a predictable texture cost.

AI USE STATEMENT

In this work, we used generative AI tools for editing the manuscript to improve readability, for formatting references and compiling the bibliography, and for checking the manuscript against the venue's formatting requirements. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims or assisting with proofs, proposing or refining hypotheses, designing the methodology or the experiments, implementing the methods, or interpreting the results; assisting with translation, cleaning and reformatting data sets, and supporting qualitative or thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work: every AI-suggested edit was accepted or rejected by an author, and every bibliographic entry added or reformatted with AI assistance was checked by an author against the original source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All diagnostic and training procedures are specified in Appendix B (hyperparameters), Appendix C (ANOVA and bootstrap protocols), and Appendix E (the full 200 H200-hour budget). The position probe, including the single-backward hook attribution and its self-test against brute-force per-bucket backwards, is released as a standalone file (`experiments/` directory of the supplementary material); the self-test verifies that hook bucket gradients equal per-bucket brute-force backwards and that they sum to the full gradient. All figures are generated by a script from a single canonical data file, so every number in the text, tables, and figures is traceable to one source.

ETHICS STATEMENT

This work is a methodological study of optimization interference in unified multimodal models and does not involve human subjects, personal data, or new data collection. All experiments use publicly released model checkpoints (Show-o and Janus-Pro) and publicly available evaluation benchmarks (MME, POPE, GenEval, and MJHQ-30k) under their respective licenses, and we release no new data set. The position-resolved interference map is a diagnostic tool and position-aware modulation is a training procedure; neither extends the generative capabilities of the underlying models beyond those of the base checkpoints, so we do not anticipate misuse risks beyond those already associated with image-generation models (e.g., the synthesis of misleading imagery), and our models inherit any biases present in the pretraining data and benchmarks of those checkpoints. The authors declare no conflicts of interest and no sponsorship that could have influenced the results. We have read and adhere to the ICLR Code of Ethics.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://arxiv.org/abs/1610.01644.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.09673.

Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2202.04200.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 2020a. URL https://proceedings.mlr.press/v119/chen20s.html.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 793–802. PMLR, 2018. URL http://proceedings.mlr.press/v80/chen18a.html.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Advances in Neural Information Processing Systems*, volume 33, 2020b. URL https://arxiv.org/abs/2010.06808.

Zisheng Chen, Chunwei Wang, Runhui Huang, Hongbin Xu, Xiuwei Chen, Jun Zhou, Jianhua Han, Hang Xu, and Xiaodan Liang. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2503.06764.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity, 2018. URL https://arxiv.org/abs/1812.02224.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2012.09841.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38, 2025. URL https://arxiv.org/abs/2306.13394.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA SEE and draw with SEED tokenizer. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01218.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023. URL https://arxiv.org/abs/2310.11513.

Yiwei Guo, Shaobin Zhuang, Zhipeng Huang, Canmiao Fu, Chen Li, Jing Lyu, and Yali Wang. Wintok: A win-win hybrid tokenizer via decomposing visual understanding and generation with transferable tokens, 2026. URL https://arxiv.org/abs/2605.18115.

Yudong Han, Yong Wang, Zaiquan Yang, Zhen Qu, Liyuan Pan, and Xiangxiang Chu. Visual enhanced depth scaling for multimodal latent reasoning, 2026. URL https://arxiv.org/abs/2604.10500.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.24365.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

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 (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.04122.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. URL https://arxiv.org/abs/2402.17245.

Yan Li, Ning Liao, Xiangyu Zhao, Shaofeng Zhang, Xiaoxing Wang, Yifan Yang, Junchi Yan, and Xue Yang. Evotok: A unified image tokenizer via residual latent evolution for visual understanding and generation, 2026. URL https://arxiv.org/abs/2603.12108.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.20.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2411.04996.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.07965.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2110.14048.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.08485.

Xiangyue Liu, Zijian Zhang, Miles Yang, Zhao Zhong, Liefeng Bo, and Ping Tan. Symbiotic-moe: Unlocking the synergy between generation and understanding, 2026. URL https://arxiv.org/abs/2604.07753.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Yishun Lu and Wes Armour. Second-order multi-level variance correction for modality competition in multimodal models, 2026. URL https://arxiv.org/abs/2605.16165.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.20321.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1912.01703.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2203.15332.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.03069.

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://arxiv.org/abs/2103.00020.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Abinav Rao and Sujan Rachuri. Do understanding and generation fight? a diagnostic study of dpo for unified multimodal models, 2026. URL https://arxiv.org/abs/2603.17044.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1810.04650.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. LMFusion: Adapting pretrained language models for multimodal generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2412.15188.

Wei Song, Yuran Wang, Zijia Song, Yadong Li, Zenan Zhou, Long Chen, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2503.14324.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. URL https://arxiv.org/abs/2104.09864.

Zhaolong Su, Wang Lu, Hao Chen, Sharon Li, and Jindong Wang. Unigame: Turning a unified multimodal model into its own adversary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2511.19413.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL https://arxiv.org/abs/2406.06525.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.02905.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.11321.

Weixing Wang, Liudvikas Zekas, Anton Hackl, Constantin Alexander Auga, Parisa Shahabinejad, Jona Otholt, Antonio Rueda-Toicen, and Gerard de Melo. Beyond accuracy: Benchmarking cross-task consistency in unified multimodal models, 2026. URL https://arxiv.org/abs/2604.25072.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/1905.12681.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.05874.

Xiwen Wei, Mustafa Munir, and Radu Marculescu. Mitigating intra- and inter-modal forgetting in continual learning of unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2512.03125.

Xiwen Wei, Mark Nutter, Madhusudhanan Srinivasan, and Radu Marculescu. Pareto lora: Mitigating modality imbalance in unified multimodal models via pareto-optimal gradient integration, 2026. URL https://arxiv.org/abs/2606.17296.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, June 2025a. doi: 10.1109/cvpr52734.2025.01210. URL http://dx.doi.org/10.1109/cvpr52734.2025.01210.

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system, 2026. URL https://arxiv.org/abs/2609.01607.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://arxiv.org/abs/2409.04429.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2408.12528.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2506.15564.

Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2209.11379.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.10789.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.05737.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2411.00776.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, Jie Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.22485.

Jiaxing Zhang and Hao Tang. Resolving task objective conflicts in unified model via task-aware mixture-of-experts. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026. URL https://arxiv.org/abs/2506.03591. Extended abstract.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.11039.

A Extended related work

Autoregressive visual generation and raster order. The premise that visual-token roles vary with position has a long history in autoregressive image modeling. Image GPT (Chen et al., 2020a) predicts pixels in raster order with a plain Transformer (Vaswani et al., 2017); discrete tokenizers (van den Oord et al., 2017; Esser et al., 2021; Yu et al., 2024) made the same recipe scalable to high-resolution text-to-image generation (Ramesh et al., 2021; Yu et al., 2022; Sun et al., 2024), and masked-token generation (Chang et al., 2022) decodes tokens in parallel over a few iterations rather than in raster order. Two recent lines make position an explicit design variable: next-scale prediction (Tian et al., 2024) orders tokens from coarse to fine so that early steps fix global structure, and randomized autoregressive training (Yu et al., 2025) anneals from random permutations to raster order and shows that the factorization order materially affects sample quality. Rotary position encoding (Su et al., 2024) is the mechanism through which the backbones we study bind a token to its raster index. Our map measures a consequence of this positional structure that these works do not: how much the gradient from each position conflicts with a second objective sharing the same parameters.

Unified models and unified tokenizers. Beyond the systems named in Section 1, unified models differ mainly in how the understanding pathway meets the visual-token stream. SEED-LLaMA (Ge et al., 2024) and VILA-U (Wu et al., 2025b) tokenize images into semantically aligned discrete codes so that a single next-token objective serves both tasks. TokenFlow (Qu et al., 2025) decouples semantic and pixel-level codebooks behind shared indices, whereas UniTok (Ma et al., 2025) argues that reconstruction and semantic supervision do not inherently conflict and enlarges the codebook instead. Show-o2 (Xie et al., 2025b) and BAGEL (Deng et al., 2025) route generation through continuous latents with flow matching; BAGEL uses a mixture-of-transformers backbone (Liang et al., 2025) whose modality-specific parameters are a form of the layer- and expert-level separation discussed in Section 2, and LMFusion (Shi et al., 2025) freezes the language modules of a pretrained LLM and trains parallel image modules, a separation motivated explicitly by forgetting of text-only capabilities. Understanding-only models trained by visual instruction tuning (Liu et al., 2023) are the reference point for the understanding axis of all of these systems. Because the position effect we measure is strongest when both objectives share the token sequence (Section 8), we expect it to be largest for the SEED and VILA-U family and smallest for designs that, like Janus-Pro, feed understanding through a separate encoder.

Multi-task optimization and modality competition. Gradient-conflict remedies for generic multi-task learning include Nash-bargaining aggregation of task gradients (Navon et al., 2022), alignment targets that pull the updates of related tasks together (Wang et al., 2021), and stochastic sign masking of inconsistent gradient components (Chen et al., 2020b); the cosine between task gradients that our map evaluates per position was proposed as an on-line signal for weighting auxiliary losses by Du et al. (2018). Large-scale re-evaluations (Kurin et al., 2022; Xin et al., 2022) find that many specialized multi-task optimizers do not beat a well-tuned scalarization, which is one reason we compare PAM against a matched-budget parameter-separation arm and a random-position control rather than against generic gradient surgery alone. The multimodal-learning literature documents that jointly trained multimodal networks can underperform their unimodal counterparts because the modalities learn at different rates (Wang et al., 2020), and corrects the imbalance by modulating each modality's gradient on the fly (Peng et al., 2022); PAM applies the same modulation idea at the finer resolution of token positions within one modality. Selecting which token positions to train on has also been shown to matter in language-model pretraining (Lin et al., 2024), which is consistent with our finding that the position structure, not the act of modulating tokens, carries the PAM effect.

Parameter-efficient adaptation and forgetting. All arms are trained with LoRA (Hu et al., 2022), which is known to forget less of the base model than full fine-tuning at the price of learning less (Biderman et al., 2024); the interaction of PAM with full fine-tuning is listed as an open question in Section 8. Optimization uses AdamW (Loshchilov & Hutter, 2019), and the probe is implemented with `torch.autograd` hooks (Paszke et al., 2019).

Frequency-domain evaluation and statistics. The band-wise FID of Table 4 decomposes images with a Laplacian pyramid (Burt & Adelson, 1983). Reporting the texture cost per frequency band follows the observation that neural networks fit low-frequency structure first (Rahaman et al., 2019), so an intervention that acts on early, low-frequency positions is expected to trade against the high-frequency band. Semantic decodability is measured with linear probes (Alain & Bengio, 2017); effect sizes are reported as partial η^2 (Cohen, 1988) and confidence intervals by paired bootstrap (Efron & Tibshirani, 1993).

B HYPERPARAMETERS AND IMPLEMENTATION DETAILS

Table 6 lists the shared configuration. The probe runs inside `torch.autograd` with forward hooks capturing module inputs and tensor hooks on module outputs; per-bucket sums are accumulated as rank-one products during the standard backward pass (Eq. 2). The map is computed on a fixed diagnostic set of 8 batches (64 understanding + 64 generation samples each) under 3 dropout seeds; 4 additional batches are held out for bucket selection and threshold calibration (Section 6), so selection never sees the reported map.

Table 6: Shared configuration.

Item	Value
Backbones	Show-o-1.3B (L=24, d=2048, T=256); Janus-Pro-7B (L=30, d=4096, T=576)
Position buckets	$K = 16$ contiguous raster buckets
LoRA	rank 16, α 32, dropout 0.05, q, k, v, o, all layers
Trainable parameters	12.6M (Show-o), 31.5M (Janus-Pro)
Optimization	AdamW, lr 10^{-4}, cosine, warmup 500, 20k steps, batch 128
Task mixture	1:1 understanding / generation
Diagnostic probe	8 batches $\times$ (64+64), 3 dropout seeds
Probe cost	$1.21\times$ (Show-o) / $1.17\times$ (Janus-Pro) vs. standard backward
PAM	$m = 6$, $\tau = 0.13$, $\mathcal{S} = \{1, \ldots, 4, 6, 7\}$ (1-indexed), $\hat{u}$ EMA refresh 2k steps
Down-weight variant	$w = 0.25$ (+6.2 GenEval pts vs. +7.7 projection)
Training overhead	$+0.3\%$ wall-clock
Seeds	3 per arm

C ANOVA AND BOOTSTRAP PROTOCOLS

The ANOVA of Table 1 treats the map cells as observations of a balanced design: each (layer, bucket) cell is observed once per module-group (5 groups: attention q, k, v, o and MLP) and per batch-seed replicate (24), yielding $L \cdot K \cdot 5 \cdot 24$ observations. Module and batch effects are additive and balanced, so the error mean square is their pooled residual; main-effect and interaction sums of squares follow the standard balanced two-way decomposition, and partial η^2 is $\mathrm{SS}_{\mathrm{effect}}/(\mathrm{SS}_{\mathrm{effect}} + \mathrm{SS}_{\mathrm{err}})$. Degrees of freedom are $\mathrm{df}_{\mathrm{layer}} = L{-}1$, $\mathrm{df}_{\mathrm{pos}} = 15$, $\mathrm{df}_{\mathrm{int}} = (L{-}1) \cdot 15$, and $\mathrm{df}_{\mathrm{err}} = L \cdot K \cdot 5 \cdot 24 - L \cdot K$.

Bootstrap CIs resample the GenEval prompt set and the MME item set (10^4 resamples) with paired assignment across arms, and the percentile interval is reported. Seed variance is the standard deviation of the arm metric across 3 seeds.

D PER-BAND POSITION PROFILES

Figure 7 plots the position profile of the conflict map separately for the shallow, middle, and deep layer bands, with band means $-0.10/-0.02/-0.135$ (Show-o) and $-0.05/-0.005/-0.095$ (Janus-Pro). The middle band is nearly flat and slightly positive at late positions; conflict is a joint

function of depth band and position, which is what the significant interaction in Table 1 quantifies ($\eta^2_{\text{int}} = 0.085/0.068$).

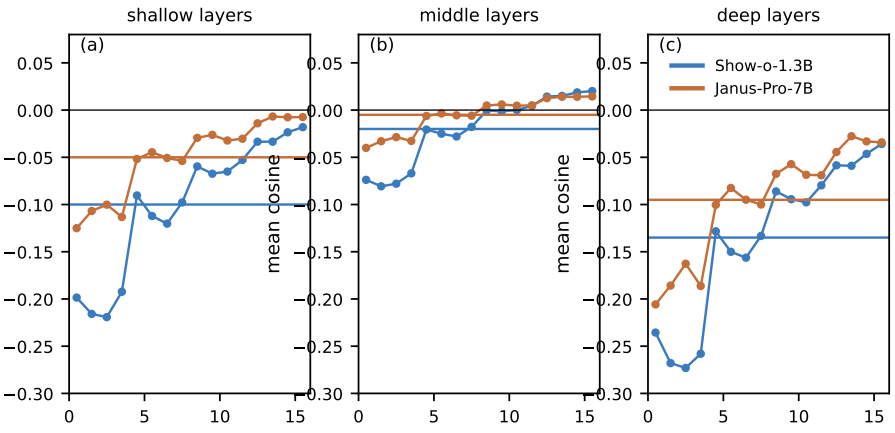


Figure 7: Position profiles of the interference map per layer band (shallow / middle / deep), both models; horizontal lines mark band means.

E Compute

Figure 8 breaks the total budget of 200 H200-hours: 12 for environment bring-up and throughput measurement, 28 for the map (E1), 8 for the norm–direction decomposition (E2, reusing E1 recordings), 10 for the semanticity attribution (E3), 78 for the four-arm matched-budget training (E4), 30 for the cross-architecture study (E5), and 34 for evaluation and re-runs.

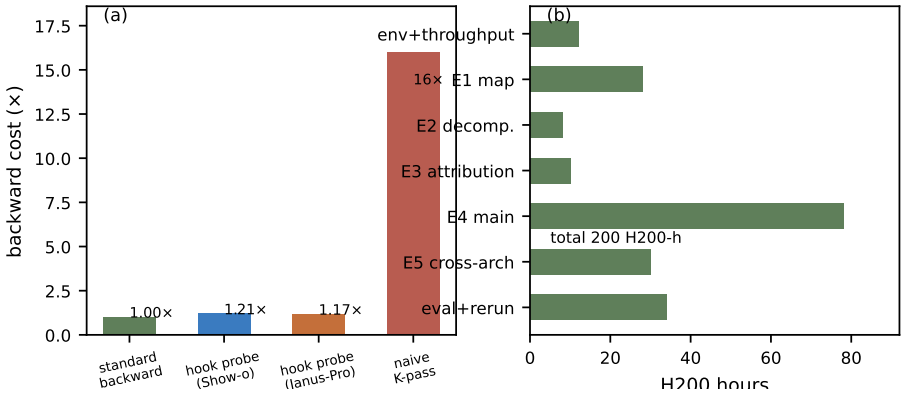


Figure 8: (a) Backward cost of the probe vs. alternatives. (b) H200-hour budget by stage; the total is 200 hours.

F Additional result figures

Figure 9 shows the frequency-band FID decomposition of Table 4, including the high-frequency cost of PAM ($+4.1\%$ relative to layer-wise separation) against its low-frequency gain (-1.0%) and unchanged overall FID. Figure 10 visualizes the cross-architecture comparison of Table 5: effect sizes by factor, per-metric deltas with CIs, and the random-control recovery contrast (31.6% vs. 60.0% on MME).

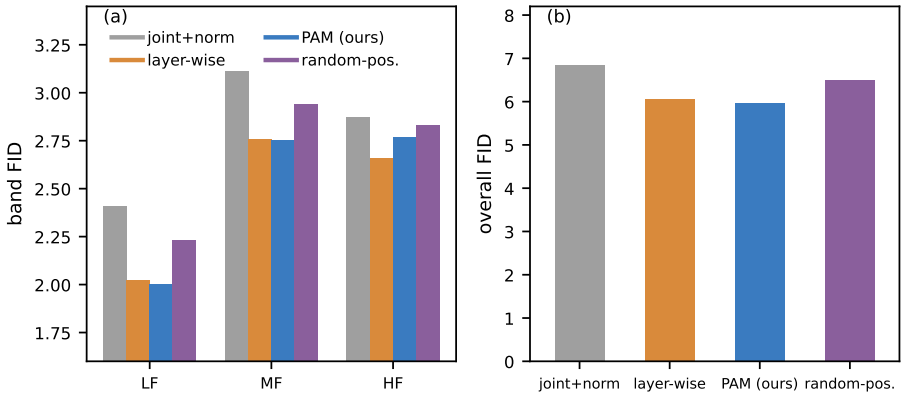


Figure 9: Frequency-band FID (a) and overall FID (b) on Show-o-1.3B for the four arms of Table 4.

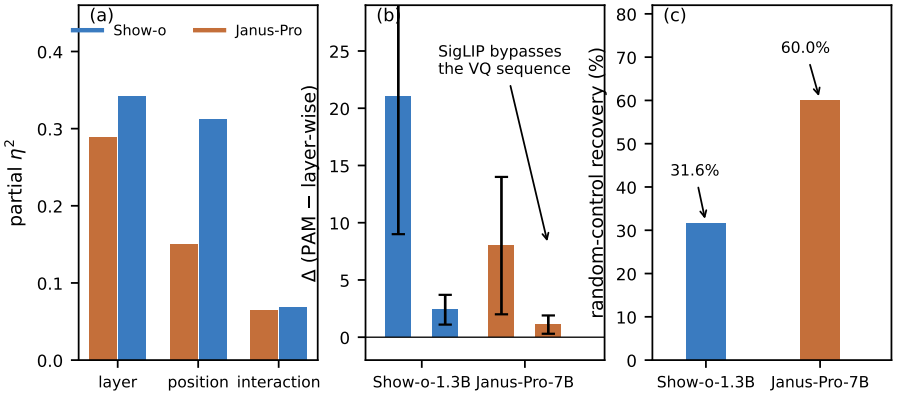


Figure 10: Cross-architecture comparison. (a) ANOVA effect sizes by factor. (b) Per-metric deltas of PAM over layer-wise separation with 95% CIs; on Janus-Pro the SigLIP understanding branch bypasses the visual token sequence and the effect halves. (c) Random-control recovery of the PAM gain (MME).