TOWARDS PHYSICS OF MULTIMODAL PRETRAINING: KNOWLEDGE FLOW, MODALITY SYNERGY, EARLY UNIFICATION, AND RECIPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision offers a critical axis for advancing foundation models, driving a shift towards natively unified multimodal pretraining. Despite this momentum, the design space and the fundamental mechanisms of how modalities interact during unified training remain underexplored. We provide empirical clarity through a systematic exploration of multimodal pretraining. Our controlled experiments on both synthetic and large-scale real-world datasets yield four key insights into the physics of multimodal pretraining: **(i) Knowledge Flow:** We disentangle how language, visual understanding, and visual generation transfer knowledge across modalities, revealing distinct patterns of influence and asymmetry; **(ii) Synergy vs. Competition:** We show that data "complexity" largely determines whether modalities are synergistic, identify architectural choices that promote synergy—such as shared attention and normalization with modality-specific feed-forward layers, and find that these behaviors generalize across different visual tokenizer designs; **(iii) Early Unification:** Unifying modalities from the very early stages and training them jointly is shown to be more effective than late alignment or sequential training. This process uncovers a vision laziness phenomenon, where delayed integration leads models to rely on language priors; **(iv) Recipes:** We derive efficient pretraining recipes that achieve strong generative performance using only 5% of the compute budget. These core findings are subsequently validated at scale by training multiple 13.5B MoE models on 2T tokens.

We strongly recommend a visit to our `project page` for an interactive reading.

1 INTRODUCTION

The trajectory of foundation models is rapidly evolving from unimodal language to multimodal (GPT4o, 2024; Gemini, 2025; Qwen3.5-omni, 2026). While earlier efforts integrated vision via late-fusion (Li et al., 2023; Liu et al., 2023a; Alayrac et al., 2022; Tong et al., 2024a; Qwen3-vl, 2025; Llama3, 2024), forcing visual representations to conform post-hoc into a pre-existing text space inherently limits cross-modal capacity. To overcome this, early-fusion unifies visual and linguistic representations from the start of pretraining, enabling them to co-evolve for richer understanding (K2.5, 2026; Llama4, 2025; Gemini, 2025; TML, 2026). Recently, this paradigm has expanded into fully unified multimodal models (Yang et al., 2026a; Deng et al., 2025; Tong et al., 2026a; Team, 2026b; Wang et al., 2026b; Wu et al., 2025b; Liu et al., 2025; Han et al., 2026a; Team, 2024; Li et al., 2025c; Wang et al., 2026c), pursuing simultaneous visual understanding and visual generation alongside language within a single network.

However, unified multimodal pretraining dramatically broadens the design space, leaving model construction guided largely by heuristics (Tong et al., 2026a). The underlying mechanisms—the "physics" (Allen-Zhu, 2024) governing unified pretraining—remain poorly understood. Most existing models are built by retrofitting pretrained LLMs/MLLMs (Shi et al., 2024; Deng et al., 2025; Xie et al., 2025; Liu et al., 2026; Diao et al., 2026), appending generative components rather than studying how vision should fundamentally shape the representation from scratch. In this work, we replace intuition with systematic empirical exploration. Across controlled synthetic benchmarks and web-scale datasets, we isolate the fundamental behaviors of multimodal interaction to establish a principled foundation for unified modeling. Our exploration yields four core insights:

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.

- **Knowledge Flow (§ 3).** Modality transfer across language, understanding, and generation is starkly asymmetric: language universally boosts all visual tasks, and visual understanding strongly seeds visual generation, whereas generation exhibits neutral backward transfer. On controlled synthetic data (Johnson et al., 2017), we further reveal that cross-modal transfer is concept-dependent: zero-shot transfer fails for low-level attributes (color, shape), but succeeds for structural concepts.
- **Modality Synergy (§ 4).** Whether modalities compete or synergize depends on data complexity and parameter allocation. Highly simplified tasks act as cross-modal catalysts, whereas complex distributions cause capacity competition. Mechanistically, sharing attention and normalization layers fosters synergy, while isolating feed-forward networks (FFNs) prevents competition. These behaviors generalize consistently across diverse visual tokenizers.
- **Early Unification (§ 5).** Joint pretraining from early stages decisively outperforms late-stage integration or sequential training curricula. We uncover a "vision laziness" phenomenon: delaying visual unification causes models to rely heavily on language priors, leading to suppressed visual activations and reduced attention on image tokens.
- **Recipes (§ 6).** Leveraging our transfer asymmetry insights, we show that visual generation can reach near-peak fidelity using only 5% of the training token budget. We scale these findings by pretraining 13.5B MoE models on 2T tokens under controlled single-variable baselines, verifying that early unification and asymmetric mixtures generalize at scale.

2 Experimental Setup

Training protocol. We pretrain decoder-only Transformers adhering to the **Llama-3** architecture (Llama3, 2024) (SwiGLU, RoPE $\theta = 500{,}000$, pre-RMSNorm, GQA, QK-norm, FlashAttention). We adopt the **Transfusion** framework (Zhou et al., 2025) unifying discrete next-token prediction for text and continuous flow matching (x-prediction) for visual generation. Our controlled baseline is a 1.5B Llama-3-like model featuring **modality-specific split FFNs** (16 layers, hidden dim 2048, 32 query heads, 8 KV heads; 2.3B total params). By default, visual tokens are encoded via frozen SigLIP-2 ViT-400m/14 (Zhai et al., 2023; Tschannen et al., 2025) into a 16×16 grid (256 tokens) and decoded via Representation Autoencoders (**RAE**) (Zheng et al., 2026). We also systematically benchmark three alternative tokenizers: **Raw Pixels**, **CLIP + VAE**, and discrete autoregressive **UniTok** (Jiao et al., 2025). Optimization uses AdamW, cosine decay, 8k warmup steps, and FSDP-2 with 4096 context length. Pretraining uses DCLM text (Li et al., 2024b) and 350M Shutterstock image-text pairs (SSTK). Following Cambrian-1 (Tong et al., 2024a), understanding models receive 1 epoch of SFT on Cambrian-7M. Full hyperparameter and architectural specifics are in Appendix D.

Evaluation protocol. We assess models across three foundational modalities: **Language:** 11 standard downstream reasoning and open-ended QA benchmarks, alongside in- and out-of-distribution validation perplexities; **Visual Understanding:** 16 VQA benchmarks spanning four primary axes: General, Knowledge, OCR & Chart, and Vision-Centric tasks; **Visual Generation:** Compositional generation fidelity, dense-prompt following, multi-bucket text-image alignment, and held-out validation diffusion loss. Full benchmark breakdowns and formal citations are provided in Appendix D.2.

3 Demystifying Modality Knowledge Flow

Unified models jointly train on language, image understanding (image $\rightarrow$ text), and image generation (text $\rightarrow$ image). It is still not fully clear how knowledge flows between these three abilities, with recent studies often reaching different conclusions (Zhang et al., 2025; Pan et al., 2025; Deng et al., 2025; Tong et al., 2026c;a; Shi et al., 2026; Niu et al., 2025a). In this section we systematically dissect these directions of transfer through controlled mixture experiments. We organize the study in two stages: first on general, web-scale data (SSTK and DCLM) (§ 3.1), and second on a procedurally generated synthetic benchmark (CLEVR) to isolate specific concept transfers causally (§ 3.2).

3.1 Modality transfer on real-world data

Finding 1: Modality knowledge flow is asymmetric: language universally boosts all visual tasks, and visual understanding strongly seeds visual generation, whereas visual generation exhibits neutral backward transfer.

3.1.1 LANGUAGE PRIORS FOR VISUAL UNDERSTANDING AND GENERATION

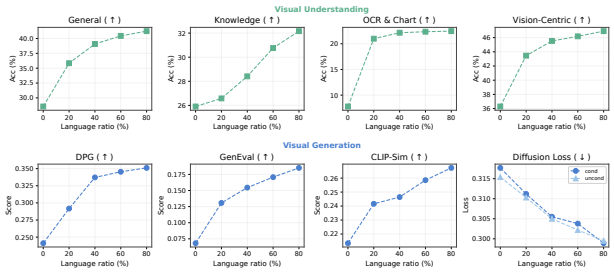


Figure 1: **Impact of scaling language data on visual understanding and generation.** Increasing the language ratio universally improves both vision capabilities.

Setting. We ask whether adding language data on top of a fixed vision allocation helps either visual generation or visual understanding. Models are trained *from scratch* on language and vision data. The vision allocation is held constant at 50B tokens, and language tokens are added on top so that the language ratio sweeps $\{0\%, 20\%, 40\%, 60\%, 80\%\}$, yielding added language tokens of $\{0, 12.5, 33, 75, 200\}$ B.

Results. As shown in Figure 1, increasing the language data ratio from 0% to 80% yields monotonic improvements across all visual understanding evaluation axes. General and Vision-Centric tasks exhibit steady improvements, while Knowledge and OCR & Chart capabilities see dramatic relative increases. For visual generation, a stronger language prior uniformly enhances text-to-image alignment and compositional generation quality. Notably, the diffusion loss decreases for both conditional and unconditional generation, indicating that joint language pretraining imparts benefits that universally improve the model's native visual modeling capacity. This aligns with recent observations of language visual priors (Han et al., 2026b; Wang et al., 2025b) and supports the Platonic Representation Hypothesis (Huh et al., 2024; Huang et al., 2025; Ziyin & Chuang, 2025), suggesting that a mature language manifold naturally captures an underlying, modality-agnostic world structure that bootstraps visual learning.

3.1.2 TRANSFER FROM VISION UNDERSTANDING TO LANGUAGE AND VISUAL GENERATION

Setting. We study the effect of adding visual understanding data on top of a fixed language or visual generation allocation. The base allocation (either language or generation) is held constant at 50B tokens, and visual understanding tokens are added on top so that the understanding ratio sweeps $\{0\%, 20\%, 40\%, 60\%, 80\%\}$, yielding added understanding tokens of $\{0, 12.5, 33, 75, 200\}$ B. When studying the effect on language, models are trained from scratch on the mixed data. When studying visual generation, however, we initialize training from a 50B pretrained language model rather than from scratch; this ensures the model possesses very basic text capabilities. As before, we evaluate both unconditional and text-conditional generation variants.

Results. Figure 2 illustrates a mild capability trade-off on the language side. As the proportion of visual understanding tokens increases, language benchmark scores experience a slight drop and perplexity marginally worsens. This degradation is largely an artifact of the text distribution in standard vision-language datasets being fundamentally different from that of pure language datasets. Conversely, on the vision side, visual understanding serves as a powerful catalyst for visual generation. Increasing the understanding data ratio markedly improves generation evaluation metrics and significantly reduces both conditional and unconditional diffusion losses. This confirms that the visual features learned during understanding tasks transfer highly effectively to generative processes.

3.1.3 DOES VISUAL GENERATION HELP UNDERSTANDING AND LANGUAGE?

Setting. We ask whether adding visual generation data on top of a fixed language or visual understanding allocation helps either capability. The base allocation (either language or understanding) is held constant at 50B tokens, and generation tokens are added on top so that the generation ratio sweeps $\{0\%, 20\%, 40\%, 60\%, 80\%\}$, yielding added generation tokens of $\{0, 12.5, 33, 75, 200\}$ B. When studying the effect on language, models are trained from scratch on the mixed data; the fixed language compute budget is jointly trained with the added generation tokens (in both unconditional and text-conditional variants). When studying visual understanding, however, we initialize training from a 50B pretrained language model to ensure foundational text abilities. The generation data is then mixed with the fixed understanding allocation, and each pretrained model is subsequently fine-tuned on Cambrian-VQA to measure downstream understanding performance.

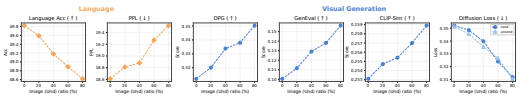


Figure 2: **Scaling visual understanding data.** Benefits visual generation but degrades pure language.

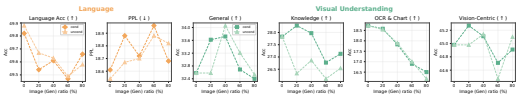


Figure 3: **Scaling visual generation data.** Causes minor fluctuations in language and understanding tasks.

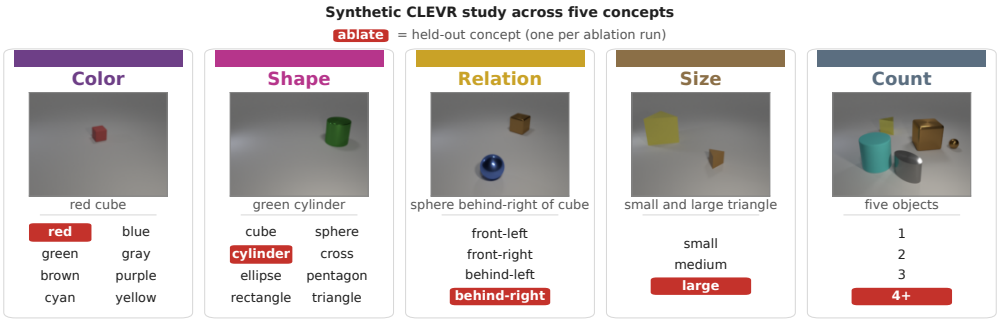


Figure 4: **Overview of the synthetic CLEVR testbed.** We extend the standard CLEVR vocabulary across five conceptual axes: color, shape, spatial relation, size, and object count. Target concepts (highlighted in red) are systematically ablated from targeted modality training streams.

Results. As depicted in Figure 3, introducing visual generation data does not demonstrate a positive transfer to other tasks. Scaling the generation ratio leads to minor fluctuations in language benchmark accuracy and perplexity, rather than a significant degradation. Similarly, for visual understanding, performance across the General, Knowledge, OCR & Chart, and Vision-Centric axes exhibits marginal fluctuations without a distinct trend. Unlike other cross-modal directions, visual generation does not act as a strong catalyst to boost language or understanding performance. However, it also does not severely interfere with them. Generative training objectives (such as flow matching) do not inherently conflict with next token prediction, though they offer limited backward knowledge transfer.

3.2 Concept Transfer: A Synthetic Controlled Study

Finding 2: Knowledge flow is concept-dependent: zero-shot transfer fails for low-level attributes (color, shape), whereas visual understanding provides zero-shot transfer for structural concepts to generation, and generation provides latent priors that accelerate low-level understanding.

3.2.1 Data Pipeline and Experiment Setting

To isolate concept transfer causally without web-scale confounders, we build a diagnostic testbed on procedurally generated CLEVR (Johnson et al., 2017) with expanded color and shape vocabularies (Figure 4). We ask: (1) Does learning a visual concept entirely via one objective transfer zero-shot to another? (2) If zero-shot transfer fails, does it impart latent priors that accelerate subsequent learning?

We generate three parallel streams per rendered scene: text-to-image targets, captions, and VQA pairs. Leveraging scene-graph metadata, we cleanly ablate targeted concepts from exactly one modality stream while keeping the other intact:

- **Understanding → Generation:** The concept is filtered from generation targets but retained in captions and VQA.
- **Generation → Understanding:** The concept is excluded from captions and VQA but retained in generation data.

We test five concept categories: color, shape, spatial relation, size, and count. Each directional ablation is benchmarked against an all-concept baseline and a zero-exposure control. Evaluation uses held-out VQA accuracy for understanding and Qwen3-VL-8B-Instruct (Qwen3-vl, 2025) as an automated judge for generated images (details in Appendix F).

3.2.2 ZERO-SHOT CONCEPT TRANSFER RESULTS

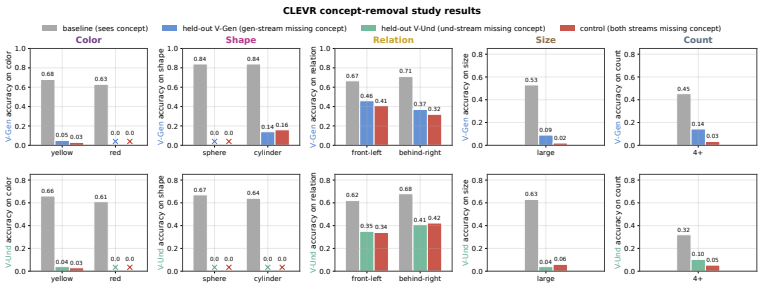


Figure 5: **Zero-shot concept transfer on CLEVR. Left two:** Low-level attributes fail in both directions. **Right three:** Structural concepts transfer asymmetrically.

Semantic attributes strictly do not transfer. For low-level concepts (color, shape), zero-shot transfer fails completely in both directions. Ablating them collapses accuracy to control levels, showing basic visual vocabularies require dedicated objectives (Wu et al., 2025b; Li et al., 2025c).

Understanding transfers to generation for structures. For spatial relations, size, and count, retaining them solely in understanding yields generation accuracy above control.

Generation offers minimal backward transfer. Conversely, learning structural concepts solely through generation does not broadly equip the model to understand them.

3.2.3 PRIORS IN CONCEPT TRANSFER

For low-level concepts (such as color and shape) that completely failed to transfer zero-shot in §3.2.2, we investigate whether any latent knowledge is retained beneath the surface. Although these concepts cannot be utilized immediately without explicit task-specific exposure, training on one stream might establish supportive representations that can be unlocked via fine-tuning.

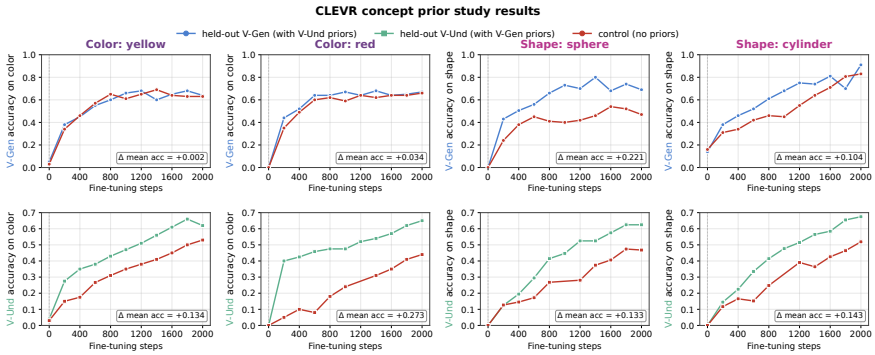


Figure 6: **Concept recovery via fine-tuning. Top:** Understanding yields minimal generative priors. **Bottom:** Generation strongly accelerates understanding recovery.

Settings. We fine-tune ablated checkpoints for 2000 steps on streams containing the held-out concept to test latent representation formation against a no-prior control.

Results. Figure 6 reveals an unexpected asymmetry: **(1) Generation provides a strong latent prior for understanding.** Prior generative exposure recovers VQA understanding significantly faster (Δ mean acc $+0.133$ to $+0.273$), as dense pixel distributions map readily to text. **(2) Understanding provides minimal priors for generation.** Prior exposure via understanding offers virtually no advantage for generating colors (Δ acc ≈ 0) and minimal gains for shape.

4 SYNERGY VS. COMPETITION IN UNIFIED PRETRAINING

A unified model allocates a single parameter set across visual generation, visual understanding, and language modeling. Whether these objectives reinforce one another (**synergy**) or compete for capacity (**competition**) determines the viability of unification. We dissect this question along three complementary axes: the data and task complexity of each modality (§ 4.1), parameter sharing between modalities (§ 4.2), and generalization across multiple vision encoder designs. The first

asks whether harder visual or linguistic data taxes the other modality more; the second asks whether weight-sharing strategies can convert latent competition into synergy; the third evaluates whether these interaction patterns remain consistent under different visual representation spaces.

Finding 3: Task complexity and parameter sharing dictate modality interaction. Simple tasks boost cross-modal learning, whereas complex tasks induce capacity competition. Sharing attention and normalization fosters synergy, while decoupling FFNs mitigates competition.

4.1 Data and task complexity

Whether two modalities cooperate or compete during unified pretraining is not solely governed by architecture; it is deeply tied to the intrinsic complexity of the data they are asked to learn. We hypothesize that a noisy, low-information visual stream imposes fundamentally different demands on a shared parameter space than a highly structured natural-language stream. Consequently, the magnitude of cross-modal interference, or the potential for synergy, predictably scales with the semantic richness and difficulty of the tasks. To systematically probe this, we isolate the task complexity of one modality while holding the other completely fixed.

Setting. We isolate the task complexity of one modality on a fixed 100B token budget (50B language / 50B vision). As illustrated in Figure 11 (Appendix D), we evaluate seven rungs of vision complexity (solid backgrounds, pattern noise, unstructured noise, text rendering, OCR images, video, SSTK images) and three rungs of language complexity (simple repeated letters, nonsense consonant strings, DCLM web-text). We report cross-modal text perplexity and validation diffusion loss.

Results. As shown in Figure 7(a), task complexity dictates the balance between competition and synergy: **Simple tasks act as cross-modal boosters.** Low-complexity tasks improve performance on the counterpart modality. Pairing text with simple backgrounds or noise improves language perplexity (Δ ppl -0.211 and -0.183) over pure language baselines. Symmetrically, simplified text yields the largest reduction in visual diffusion loss ($\Delta - 0.0168$). **Complex tasks induce capacity competition.** Escalating the visual stream to high-entropy distributions (video, SSTK) degrades text perplexity (Δ ppl $+0.052$ and $+0.075$). As semantic entropy rises, capacity competition eventually outweighs cross-modal synergy.

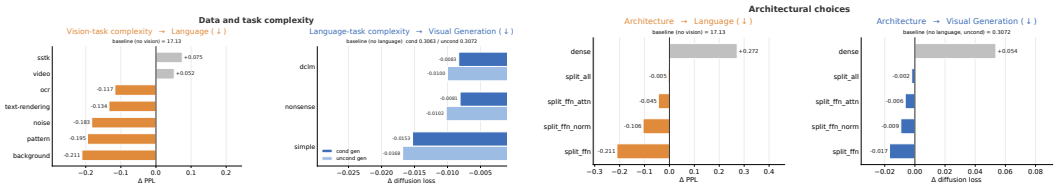
These results reveal that unified pretraining is defined by a delicate interplay of two opposing forces: cross-modal synergy and capacity competition. We next investigate where these synergies and competition manifest within the model's internal components.

4.2 Parameter-sharing in Transformers for promoting synergy

To understand how the architecture of the network dictates these modality interactions, we dissect the standard Transformer block. Specifically, we evaluate which components require specialization to mitigate capacity competition, and which must remain shared to foster cross-modal synergy.

Setting. We freeze the data mixture at a 50%/50% split (DCLM and SSTK) and deconstruct the Transformer block into three configurable parameter groups: FFN, Attention (Q/K/V/O), and FinalNorm, testing five configurations: `dense` (all shared), `split_ffn` (FFNs decoupled per modality), `split_ffn_attn` (FFNs and Attention decoupled), `split_ffn_norm` (FFNs and Normalizations decoupled), and `split_all` (all isolated).

Results. As depicted in Figure 7(b), internal parameter allocation governs modality interactions: **Dense networks suffer from severe competition.** The `dense` model degrades language ppl by $+0.272$ and vision loss by $+0.0537$ compared to unimodal baselines. **Decoupling FFNs mitigates competition.** Isolating feed-forward networks (`split_ffn`) eliminates capacity competition while preserving cross-modal gains (language ppl -0.211, vision loss -0.0168). This strongly supports MoE/MoT architectures (Liang et al., 2024; Lin et al., 2024; Team, 2026b; Tong et al., 2026a). **Shared attention and norm drive synergy.** Decoupling attention (`split_ffn_attn`) or normalization (`split_ffn_norm`) drastically shrinks improvements, and isolating all parameters (`split_all`) collapses to unimodal baseline levels. Shared attention aligns representations, while split FFNs absorb modality-specific computational demands.



(a) **Impact of task complexity.** Visual complexity turns synergy into competition (left); simple language boosts generation (right).

(b) **Parameter sharing choices.** Fully shared parameters force competition; `split_ffn` mitigates it while preserving synergy.

Figure 7: **Sources of synergy and competition in unified pretraining.** (a) Task complexity dictates cross-modal interactions (100B token budget; see sample images in Appendix D). (b) Deconstructing the Transformer block: decoupling FFNs isolates capacity competition, while shared attention and norm preserve cross-modal synergy.

Whether these interaction patterns remain consistent across visual representation spaces is also verified; as detailed in Appendix H, these synergy behaviors and asymmetric transfers consistently generalize across four distinct vision tokenization designs.

5 The Necessity of Early Unification

Finding 4: Early and simultaneous unification is essential. Delaying visual integration consistently impairs native visual capabilities, while sequential training induces catastrophic forgetting and fails to unlock cross-modal synergy.

Given that vision heavily benefits from language priors, a natural intuition might be to pretrain a pure language model first and align visual capabilities later. In this section, we challenge this common practice. By systematically varying the timing of visual integration (§ 5.1) and the sequencing (§ 5.2) of modality training, we investigate the underlying temporal mechanics of multimodal learning. Our findings challenge the prevailing late-alignment heuristics, revealing why early and simultaneous unification is a necessity to prevent vision laziness and unlock more cross-modal synergy (§ 5.3).

5.1 Early unification vs. late unification

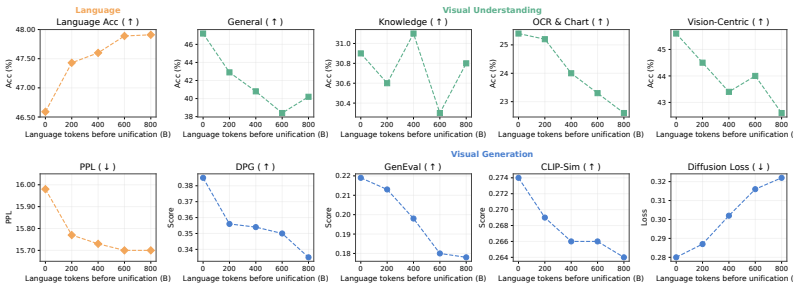


Figure 8: **Timing of unification training.** Extending language pretraining yields diminishing text returns while triggering a steep decline across vision.

Setting. Holding total compute fixed at 1T tokens, we sweep pure-language tokens before unification over $\{0, 200, 400, 600, 800\}$B, spending the remainder in unified training (50/50 language/vision mix). The 0B point is the unified-from-scratch baseline.

Results. Figure 8 shows that while extending initial language training initially improves text metrics, these gains plateau between 600B and 800B tokens. In stark contrast, visual understanding (General, OCR, Vision-Centric) and all generative metrics suffer a steep, consistent decline. Pushing unification late into training systematically cripples visual capabilities.

5.2 Sequential vs. joint training

Late-fusion style training often implicitly assumes a language-first progression, with vision introduced only afterward. More broadly, this raises a natural question: since humans typically acquire vision before language (Steinberg & Steinberg, 1975; Smith & Gasser, 2005; Orhan & Lake, 2024; Vong et al., 2024; Bambach et al., 2018), would a vision-first training order be more effective than a

language-first one? Beyond studying unified timing, we therefore examine whether multimodal learning can be decomposed into discrete sequential stages.

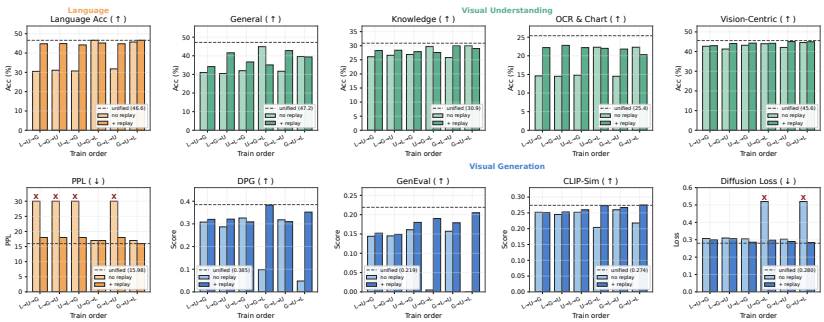


Figure 9: **Impact of sequential vs. joint pretraining.** Joint training decisively dominates all sequential approaches.

Settings. Holding total budget at 1T tokens (50/25/25 L/U/G split), we evaluate all six sequential modality orderings, with and without a 12.5% replay buffer, against simultaneous joint training.

Results. As shown in Figure 9, joint training decisively outperforms every sequential ordering across virtually all metrics. Even with replay to mitigate catastrophic forgetting (Wu et al., 2025c; Zhai et al., 2024), sequential training fails to foster active co-evolution.

5.3 Vision Laziness in Late Alignment

Finding 5: Late alignment causes vision laziness: longer language warmups progressively suppress visual activations, shrink wrapper embeddings, and reduce visual attention percentage.

The previous studies established that delaying multimodal unification systematically degrades visual performance. Here we ask *why*: what changes inside the model when the language trunk is allowed to harden before vision is introduced? We probe this question along four axes, and find a single coherent mechanism: the vision pathway becomes increasingly underdeveloped and disconnected from the language manifold the more pretrained the language trunk is at the moment of unification.

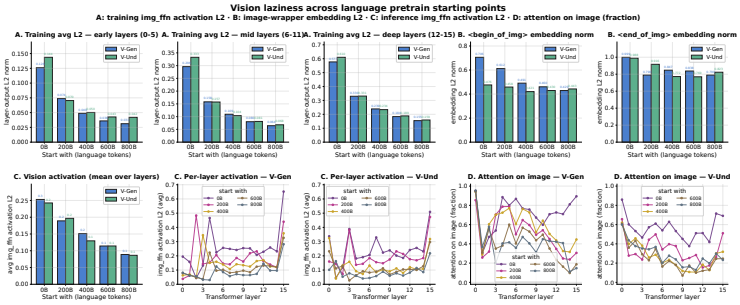


Figure 10: **Vision laziness in late alignment.** Longer pure-language warmup suppresses visual activations, reduces wrapper embeddings, and drops attention on image tokens.

Setting. Initializing from language warmups of $\{0, 200, 400, 600, 800\}$B tokens followed by matched 200B joint tuning (50/50 mix), we probe the visual feed-forward branch (`img_ffn`) along four axes: **A:** Training-time activation L_2; **B:** Wrapper embedding L_2; **C:** Inference-time activation; **D:** Inference attention on visual tokens.

Results. As shown in Figure 10, delaying unification systematically induces vision laziness across all four axes: (1) Visual branches become progressively quieter during both training (A) and inference (C), contributing far less to the joint representations. (2) Wrapper token norms shrink (B), showing that the hardened language trunk resists accommodating the visual modality. (3) Attention allocated to image tokens collapses (D); instead of inspecting visual features, late-aligned models default to surrounding text shortcuts and over-rely on language priors.

6 Designing Unified Pretraining Recipes

In this section, we synthesize our empirical findings into actionable, high-performance recipes for unified multimodal pretraining. Our prior analysis of knowledge flow (§ 3) reveals a clear asymmetry:

Model	Language		Visual Understanding					Visual Generation			
	PPL ↓	Acc ↑	Gen ↑	Know ↑	OCR ↑	V-Ctr ↑	Avg ↑	DPG ↑	GenEval ↑	CLIP-Sim ↑	DiffLoss ↓
Balanced Recipe	11.97	52.86	51.50	38.90	25.15	50.14	41.42	0.676	0.467	0.310	**0.261**
Dense Model	12.14	52.03	50.12	36.66	25.43	49.74	40.49	0.667	0.459	0.308	0.266
Late-Fusion	12.25	51.78	49.89	37.03	26.22	49.50	40.66	0.672	0.471	0.308	0.269
Full	**11.67**	**54.31**	**53.63**	**40.11**	**27.23**	**51.33**	**43.08**	**0.689**	**0.482**	**0.312**	0.272

Table 1: **Scaling results and controlled baseline comparisons.** We evaluate our model against three controlled baselines to validate our main design choices: data mixture recipes (Balanced Recipe), architecture design style (Dense Model), and vision alignment strategy (Late-Fusion).

language serves as a universal booster, and visual understanding acts as a strong prior for generation, but generation offers minimal backward transfer. This implies that an optimal recipe should allocate the vast majority of the compute budget to language and understanding to efficiently "drive" generative capabilities. We first validate this asymmetric data mixture hypothesis (§ 6.1), and then integrate all our architecture and dynamic insights to validate these recipes at scale (§ 6.2).

6.1 Optimizing data mixing ratios

Finding 6: Visual generation is highly data-efficient when co-trained with dominant language and understanding priors. Highly skewed data mixtures achieve near-peak generative fidelity while preserving other capacities.

Extensive data mixing grid search. To validate whether dominant language and understanding priors can effectively drive generation, we execute an extensive grid search of mixing ratios across 1T token runs spanning three exploration axes: sweeping language proportion from 10% to 90%, varying understanding versus generation at 50% language, and exploring visual splits at 70% language. Full numerical breakdowns of all 21 results are detailed in Appendix B.

The searches consistently show that language requires the dominant compute share to preserve foundational reasoning, while visual understanding peaks around 25%. Crucially, visual generation achieves near-peak performance with only 5% token allocation: under the optimal **L70/U25/G5** split, General VQA reaches 48.3 while generative alignment scores peak (DPG 0.450, GenEval 0.237). Because visual generation offers minimal backward transfer (§ 3), skewing compute toward language and understanding establishes the rich priors that drive generation.

6.2 Scaling Recipes

To validate our core insights at scale, we pretrain 13.5B MoE models (1.5B active per token, 256 experts, top-16 routing) on a 2T token budget under controlled, single-variable ablations (Table 1).

Knowledge flow. Guided by the asymmetric transfer finding, we evaluate the skewed mixture (L70/U25/G5) against a balanced baseline (L50/U25/G25). The skewed mix boosts both language (54.31% vs. 52.86%) and understanding (43.08% vs. 41.42%), while matching generative quality (0.482 GenEval / 5.23 FID vs. 0.467 / 5.13) despite consuming $5\times$ fewer generative tokens.

Architecture. Decoupling FFNs within the 13.5B MoE consistently outperforms the 3.5B dense baseline, validating our shared-attention and split-FFN design for cross-modal synergy.

Unification timing. Early unification decisively surpasses the late-fusion counterpart (vision introduced after 60% training) across language (52.86% vs. 51.78%), understanding (41.42% vs. 40.66%), and generation loss (0.261 vs. 0.269), confirming that early multimodal interaction is essential.

7 Conclusion

Across synthetic and real distributions, we systematically uncover the mechanics of unified multimodal pretraining. We demonstrate asymmetric, concept-dependent knowledge flow, identify architectural decoupling to foster synergy, prove the necessity of early unification to prevent vision laziness, and show visual generation bootstrapped with just 5% compute. These insights offer a principled foundation for scaling natively unified foundation models.

AI USE STATEMENT

LLMs were used as assistive tools for coding, paper copy-editing, proofreading, and improving grammatical clarity of the manuscript. All scientific hypotheses, empirical experiment designs, and interpretations of results were conceived, conducted, and validated by the authors.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our empirical findings, all architectural dimensions, loss functions, and the four visual tokenization configurations are detailed in Section 2 and Appendix D. Full optimization hyperparameters (including learning rate schedules, AdamW parameters, loss weighting factors, and FSDP-2 distributed settings), along with the Cambrian-7M instruction fine-tuning protocol, are described in Appendix D. The complete procedural dataset generation, scene-graph ablations, and evaluation pipelines for our synthetic CLEVR testbed are specified in Appendix F. The evaluation protocol spanning the 11 language benchmarks, 16 visual understanding benchmarks across four axes, and visual generation metrics is documented with benchmark citations in Appendix D.2.

ETHICS STATEMENT

Our investigation is centered on fundamental modeling mechanisms and scaling dynamics using publicly available datasets (DCLM, Cambrian-7M) alongside commercially licensed assets. We did not conduct studies involving human subjects, crowdsourced human evaluations, or private/personally identifiable data. While our findings facilitate more compute-efficient pretraining recipes for generative multimodal models, capable text-to-image synthesis models carry potential risks regarding the generation of misleading visual content, hallucinated media, or societal biases reflecting the training data distributions. We believe our study provides transparent scientific insights into modality interactions and vision representations, contributing to the safe, controllable, and robust development of unified foundation models.

REFERENCES

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Zeyuan Allen-Zhu. ICML 2024 Tutorial: Physics of Language Models, July 2024. Project page: https://physics.allen-zhu.com/.

Ruichuan An, Sihan Yang, Renrui Zhang, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *Advances in Neural Information Processing Systems*, 38: 144638–144664, 2026.

Zheng Anlin, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive generation. *Advances in Neural Information Processing Systems*, 38:62656–62675, 2026.

Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. *Advances in neural information processing systems*, 31, 2018.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025a.

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. *arXiv preprint arXiv:2501.17811*, 2025b.

Yiming Chen, Junlin Han, Tianyi Bai, Shengbang Tong, Filippos Kokkinos, and Philip Torr. From pixels to feelings: Aligning mllms with human cognitive perception of images. *arXiv preprint arXiv:2511.22805*, 2025c.

Yitong Chen, Zijie Diao, Junke Wang, Lingyu Kong, Yixuan Ren, Bo He, Yu-Gang Jiang, and Zuxuan Wu. Ideal: In-depth alignment makes a discrete representation autoencoder. *arXiv preprint arXiv:2606.11096*, 2026.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

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. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture, 2026. URL https://arxiv.org/abs/2605.12500.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *ICCV*, 2025.

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 *NeurIPS D&B Track*, 2025.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy, 2026. URL https://arxiv.org/abs/2605.18678.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Gemini. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

GPT4o. gpt4o, 2024.

Ming Gui, Johannes Schusterbauer, Timy Phan, Felix Krause, Josh Susskind, Miguel Angel Bautista, and Björn Ommer. Adapting self-supervised representations as a latent space for efficient generation. *arXiv preprint arXiv:2510.14630*, 2025.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *Advances in Neural Information Processing Systems*, 38:158430–158459, 2026a.

Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pretraining. In *ICLR*, 2026b.

Xiaochuang Han, Youssef Emad, Melissa Hall, John Nguyen, Karthik Padthe, Liam Robbins, Amir Bar, Delong Chen, Michal Drozdzal, Maha Elbayad, et al. Tv2tv: A unified framework for interleaved language and video generation. *arXiv preprint arXiv:2512.05103*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, et al. Bagelvla: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*, 2026.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704, 2025.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, et al. Pi0.7: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.

Mingkai Jia, Mingxiao Li, Zhijian Shu, Anlin Zheng, Liaoyuan Fan, Jiaxin Guo, Tianxing Shi, Dongyue Lu, Zeming Li, Xiaoyang Guo, et al. Dino-tok: Adapting dino for visual tokenizers. *arXiv preprint arXiv:2511.20565*, 2025.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *CVPR*, 2025.

Jiachun Jin, Zetong Zhou, Xiao Yang, Hao Zhang, Pengfei Liu, Jun Zhu, and Zhijie Deng. Latentum: Unleashing the potential of interleaved cross-modal reasoning via a latent-space unified model. *arXiv preprint arXiv:2604.02097*, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

K2.5. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Bo Li, Yida Yin, Wenhao Chai, Xingyu Fu, and Zhuang Liu. Ueval: A benchmark for unified multimodal generation. *arXiv preprint arXiv:2601.22155*, 2026a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024a.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025a.

Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *CVPR*, 2025b.

Hongxiang Li, Hongxu Chen, Chenyang Zhu, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Spar: Semantic-pixel self-alignment and adaptive routing for unified multimodal models, 2026b. URL https://arxiv.org/abs/2606.23041.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2024b.

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 *ICML*, 2023.

Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, Jialing Tong, Haoxuan You, Xianzhi Du, Zhe Gan, Hyunjik Kim, Chao Jia, Zhenbang Wang, Yinfei Yang, Mingfei Gao, Zi-Yi Dou, Wenze Hu, Chang Gao, Dongxu Li, Philipp Dufter, Zirui Wang, Guoli Yin, Zhengdong Zhang, Chen Chen, Yang Zhao, Ruoming Pang, and Zhifeng Chen. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer, 2025c.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025d.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation, 2025.

Bin Lin, Zheyuan Liu, Chenguo Lin, Sixiang Chen, Yunyang Ge, Yunlong Lin, Jianwei Zhang, Miles Yang, Zhao Zhong, Liefeng Bo, and Li Yuan. Gear: Guided end-to-end autoregression for image synthesis, 2026. URL https://arxiv.org/abs/2606.32039.

Han Lin, Xichen Pan, Ziqi Huang, Ji Hou, Jialiang Wang, Weifeng Chen, Zecheng He, Felix Juefei-Xu, Junzhe Sun, Zhipeng Fan, et al. Exploring mllm-diffusion information transfer with metacanvas. *arXiv preprint arXiv:2512.11464*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, et al. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026.

Llama3. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Llama4. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022a.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision, 2025. URL https://arxiv.org/abs/2509.14476.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2023.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

John Nguyen, Marton Havasi, Tariq Berrada, Luke Zettlemoyer, and Ricky TQ Chen. Oneflow: Concurrent mixed-modal and interleaved generation with edit flows. *arXiv preprint arXiv:2510.03506*, 2025.

Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025a.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025b.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

A Emin Orhan and Brenden M Lake. Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*, 6(3):271–283, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Xichen Pan, Aashu Singh, Satya Narayan Shukla, Xiangjun Fan, Shlok Kumar Mishra, and Saining Xie. Repfusion: Leveraging multimodal priors for denoising in representation space, 2026. URL https://arxiv.org/abs/2606.14700.

Wujian Peng, Lingchen Meng, Yuxuan Cai, Xianwei Zhuang, Yuhuan Yang, Rongyao Fang, Chenfei Wu, Junyang Lin, Zuxuan Wu, and Shuai Bai. Unified multimodal autoregressive modeling with shared context — visual tokenizer is key to unification. In *ICML*, 2026.

Qwen3-vl. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Qwen3.5-omni. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Christian Schlarmann, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Fuselip: Multimodal embeddings via early fusion of discrete tokens. *arXiv preprint arXiv:2506.03096*, 2025.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, et al. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22488–22497, 2026.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *ICCV*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

Danny D Steinberg and Miho T Steinberg. Reading before speaking. *Visible Language*, 9(3), 1975.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Cosmos 3 Team. Cosmos 3: Omnimodal world models for physical ai, 2026a. URL https://arxiv.org/abs/2606.02800.

Meituan LongCat Team. Longcat-next: Lexicalizing modalities as discrete tokens, 2026b. URL https://arxiv.org/abs/2603.27538.

TML. Interaction models: A scalable approach to human-ai collaboration. *Thinking Machines Lab*, May 2026. doi: 10.64434/tml.20260511. https://thinkingmachines.ai/blog/interaction-models/.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024b.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2025.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining. *ICML*, 2026a.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint*, 2026b.

Yujun Tong, Dongliang Chang, Zijin Yin, Xintong Liu, Yuanchen Fang, and Zhanyu Ma. Reversing the flow: Generation-to-understanding synergy in large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6976–6986, 2026c.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Chenlong Wang, Yuhang Chen, Zhihan Hu, Dongping Chen, Wenhu Chen, Sarah Wiegreffe, and Tianyi Zhou. Quantifying the gap between understanding and generation within unified multimodal models. *arXiv preprint arXiv:2602.02140*, 2026a.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025a.

Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, et al. Ernie 5.0 technical report. *arXiv preprint arXiv:2602.04705*, 2026b.

Junke Wang, Xiao Wang, Jiacheng Pan, Xuefeng Hu, Feng Li, Jingxiang Sun, Chaorui Deng, Zilong Chen, Yunpeng Chen, Kaibin Tian, et al. Arm: An autoregressive large multimodal model with unified discrete representations. *arXiv preprint arXiv:2606.11188*, 2026c.

Luting Wang, Yang Zhao, Zijian Zhang, Jiashi Feng, Si Liu, and Bingyi Kang. Image understanding makes for a good tokenizer for image generation. *Advances in Neural Information Processing Systems*, 37:31015–31035, 2024a.

Shuai Wang, Liang Li, Yang Chen, Ruopeng Gao, Yao Teng, and Limin Wang. Uniddt: Unifying multimodal understanding and generation with decoupled diffusion transformer, 2026d. URL https://arxiv.org/abs/2606.16255.

Sophie L Wang, Phillip Isola, and Brian Cheung. Words that make language models perceive. *arXiv preprint arXiv:2510.02425*, 2025b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, Youliang Zhang, Haoqi Fan, and Xihui Liu. Representation forcing for bottleneck-free unified multimodal models. *arXiv preprint arXiv:2604.21921*, 2026e.

Zeyu Wang, Zilong Chen, Chenhui Gou, Feng Li, Chaorui Deng, Deyao Zhu, Kunchang Li, Weihao Yu, Haoqin Tu, Haoqi Fan, et al. Lightbagel: A light-weighted, double fusion framework for unified multimodal understanding and generation. *arXiv preprint arXiv:2510.22946*, 2025c.

Zirui Wang, Junyi Zhang, Jiaxin Ge, Long Lian, Letian Fu, Lisa Dunlap, Ken Goldberg, XuDong Wang, Ion Stoica, David M Chan, et al. Visgym: Diverse, customizable, scalable environments for multimodal agents. *arXiv preprint arXiv:2601.16973*, 2026f.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025b.

Haoyuan Wu, Aoqi Wu, Hai Wang, Jiajia Wu, Jinxiang Ou, and Bei Yu. Scaling native multimodal pre-training from scratch, 2026. URL https://arxiv.org/abs/2607.22043.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, et al. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*, 8, 2025c.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

xAI. grok, 2024.

Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo. In *NeurIPS*, 2025.

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 *ICLR*, 2025.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Junzhe Xu, Yuyang Yin, and Xi Chen. Tbac-uniimage: Unified understanding and generation by ladder-side diffusion tuning. *arXiv preprint arXiv:2508.08098*, 2025.

Ceyuan Yang, Zhijie Lin, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Chaorui Deng, Kunchang Li, Zihan Ding, Yuwei Guo, et al. Context unrolling in omni models. *arXiv preprint arXiv:2604.21921*, 2026a.

Cheng Yang, Chufan Shi, Bo Shui, Yaokang Wu, Muzi Tao, Huijuan Wang, Ivan Yee Lee, Yong Liu, Xuezhe Ma, and Taylor Berg-Kirkpatrick. From reasoning to pixels: Benchmarking the alignment gap in unified multimodal models. *arXiv preprint arXiv:2602.08336*, 2026b.

Hu Yu, Hang Xu, Jie Huang, Zeyue Xue, Haoyang Huang, Nan Duan, and Feng Zhao. Rae-ar: Taming autoregressive models with representation autoencoders. *arXiv preprint arXiv:2604.01545*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, KunPeng Du, Yi Wang, Limin Wang, and Yali Wang. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. In *ICLR*, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. In *CPAL*, 2024.

Guozhen Zhang, Xuerui Qiu, Yutao Cui, Tianhui Song, Changlin Li, Junzhe Li, Tao Huang, Xiao Zhang, Yang Li, Jianbing Wu, et al. Hydra-x: Native unified multimodal models with holistic visual tokenizers. *arXiv preprint arXiv:2606.13289*, 2026a.

Huichao Zhang, Liao Qu, Yiheng Liu, Hang Chen, Yangyang Song, Yongsheng Dong, Shikun Sun, Xian Li, Xu Wang, Yi Jiang, et al. Nextflow: Unified sequential modeling activates multimodal understanding and generation. *arXiv preprint arXiv:2601.02204*, 2026b.

Jihai Zhang, Tianle Li, Linjie Li, Zhengyuan Yang, and Yu Cheng. Are unified vision-language models necessary: Generalization across understanding and generation. *arXiv preprint arXiv:2505.23043*, 2025.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025.

Anlin Zheng, Haochen Wang, Yucheng Zhao, Weipeng Deng, Tiancai Wang, Xiangyu Zhang, and Xiaojuan Qi. Hita: Holistic tokenizer for autoregressive image generation. *arXiv preprint arXiv:2507.02358*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.

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 *ICLR*, 2025.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025.

A Related Work

A.1 Unified models

Multimodal models have rapidly advanced beyond treating vision simply as a conditional input for text models (Li et al., 2024a; Shukor et al., 2025; Wu et al., 2026). Instead, current research actively pursues unified architectures capable of simultaneous visual comprehension and generation (Zhang et al., 2025; Xiao et al., 2025; Geng et al., 2025; Xu et al., 2025; Xin et al., 2025; Li et al., 2025a; Wang et al., 2025c; Wei et al., 2025; Wang et al., 2025a; Li et al., 2025d; Nguyen et al., 2025; Pan et al., 2025; Li et al., 2025b; Zhang et al., 2026b; Dai et al., 2023; Sun et al., 2024; Ge et al., 2024; Dong et al., 2024; Tong et al., 2025; Pan et al., 2025; Han et al., 2025; Fu et al., 2026; Chen et al., 2025a; Wang et al., 2026e;d; Lin et al., 2026; 2025). Early unified paradigms relied on quantizing images into discrete visual tokens (Oord et al., 2017; Razavi et al., 2019; Lee et al., 2022) for autoregressive modeling (Lu et al., 2022a; Aghajanyan et al., 2022; Team, 2024; Lu et al., 2024; Wang et al., 2024b; Schlarmann et al., 2025). More recently, hybrid methods such as Transfusion (Zhou et al., 2025; Shi et al., 2024; Tong et al., 2026a) have successfully combined continuous diffusion mechanisms with discrete language prediction, establishing a highly effective standard for joint modality modeling. In parallel, advancements in unified vision representation spaces have also emerged to bridge perception and generation (Tong et al., 2026b; Zheng et al., 2026; Han et al., 2026a; Li et al., 2025c; Team, 2026b; Yue et al., 2026; Jiao et al., 2025; Lu et al., 2025; Liu et al., 2026; Anlin et al., 2026; Zheng et al., 2025; Gui et al., 2025; Jia et al., 2025; Chen et al., 2026; Zhao et al., 2025; Wang et al., 2024a; Pan et al., 2026; Singh et al., 2026; Jin et al., 2026; Zhang et al., 2026a; Yu et al., 2026; Peng et al., 2026; Li et al., 2026b; Wu et al., 2024b; An et al., 2026). Building on these advancements, our work delves into the physics of multimodal pretraining, hoping to pave the way for the design and scaling of future unified models.

A.2 Vision integration

A prevailing methodology for constructing multimodal systems relies on late-fusion. This involves stitching together pretrained LLMs (Touvron et al., 2023; Llama3, 2024) with independently pretrained visual encoders for visual understanding (Li et al., 2023; Liu et al., 2023a; Alayrac et al., 2022; Qwen3-vl, 2025) via adapters and post-hoc fine-tuning. Because rich visual signals are forced to conform to a pre-existing language space, this approach inherently bottlenecks capabilities (Tong et al., 2024a). To overcome this, the field has increasingly shifted toward early-fusion, integrating vision from the very beginning of training to allow visual and language representations to co-evolve. These early-fusion efforts focused primarily on deep, native visual understanding (K2.5, 2026; Llama4, 2025; Gemini, 2025; TML, 2026). The paradigm has expanded beyond visual understanding towards visual generation as well. By jointly modeling perception and creation, this unification enables the system to evolve towards world models (Tong et al., 2026a). This deeper synergy unlocks stronger visual interactive capabilities, which are essential for seamlessly supporting complex, multi-modal reasoning like interleaved generation (Niu et al., 2025b; Yang et al., 2026a) and real-world tasks like robotics (Hu et al., 2026; Intelligence et al., 2026; Team, 2026a). In this work, we explore the underlying physics of multimodal pretraining by systematically investigating unification timing and modality sequences, demonstrating that early unification prevents vision laziness and encourages stronger visual learning.

A.3 Knowledge flow

Recent benchmarking observations highlight that unified multimodal models often surpass specialized text-to-image systems on complex generation tasks (Niu et al., 2025b; Li et al., 2026a; Wang et al., 2026a; Yang et al., 2026b; Wang et al., 2026f), hinting that language and understanding capabilities fundamentally benefit visual generation (Tong et al., 2025; Liao et al., 2025; Han et al., 2026b). We formalize this phenomenon by rigorously dissecting the knowledge flow between modalities using both web-scale real-world distributions and strictly controlled synthetic environments. At the high level, we show that capability transfer is highly asymmetric: language acts as a universal booster, and visual understanding heavily drives generation, whereas visual generation offers minimal backward transfer. Furthermore, our controlled synthetic study uncovers that this cross-modal knowledge flow is deeply concept-dependent. Guided by these findings, which show that generative modeling

	Mix %			Language		Visual Understanding					Visual Generation			
	L	U	G	PPL ↓	Acc ↑	Gen ↑	Know ↑	OCR ↑	V-Ctr ↑	Avg ↑	DPG ↑	GenEval ↑	CLIP-Sim ↑	DiffLoss ↓
Fix MM	10	45	45	19.27	41.89	37.9	26.9	22.9	42.9	32.7	0.326	0.148	0.256	0.2984
	20	40	40	17.51	44.46	45.7	30.8	24.0	44.0	36.1	0.395	0.189	0.273	0.2802
	30	35	35	16.73	45.03	43.1	29.9	24.9	42.2	35.0	0.361	0.183	0.273	0.2826
	40	30	30	16.31	45.79	46.1	31.2	24.2	45.6	36.8	0.331	0.186	0.269	0.2946
	50	25	25	15.98	46.59	47.2	30.9	25.4	44.6	37.0	0.385	0.219	0.274	0.2804
	60	20	20	15.81	46.70	44.6	**33.0**	24.8	43.9	36.6	0.387	0.203	0.275	0.2883
	70	15	15	15.68	46.99	48.1	32.3	25.2	46.6	38.1	0.399	0.219	0.273	0.2996
	80	10	10	15.57	46.85	45.9	32.8	24.0	45.2	37.0	0.388	0.216	0.271	0.2868
	90	5	5	**15.48**	**48.08**	43.7	31.8	21.4	46.3	35.8	0.336	0.204	0.273	0.2894
Fix Lan	50	5	45	16.05	45.26	43.8	29.4	23.7	43.1	35.0	0.358	0.206	0.273	**0.2785**
	50	10	40	16.03	46.18	45.9	31.0	24.2	45.6	36.7	0.401	0.221	**0.281**	0.2801
	50	15	35	16.06	46.42	46.6	32.1	23.4	45.4	36.9	0.370	0.193	0.273	0.2834
	50	20	30	16.02	46.17	46.8	31.2	23.5	45.2	36.7	0.390	0.199	0.274	0.2909
	50	25	25	15.98	46.59	47.2	30.9	25.4	44.6	37.0	0.385	0.219	0.274	0.2804
	50	30	20	16.01	46.05	45.5	31.1	25.2	44.7	36.6	0.370	0.208	0.273	0.2893
	50	35	15	16.05	45.85	47.0	32.5	23.8	46.0	37.3	0.399	0.199	0.275	0.2821
	50	40	10	16.00	46.14	47.7	32.9	**26.5**	45.7	38.2	0.420	0.216	0.276	0.2874
	50	45	5	16.03	46.14	46.9	32.3	25.8	46.1	37.8	0.392	0.200	0.269	0.2931
Next	70	5	25	15.67	46.55	47.0	32.9	23.9	43.5	36.8	0.375	0.204	0.271	0.2823
	70	10	20	15.71	46.34	46.8	30.6	22.7	44.3	36.1	0.358	0.206	0.273	0.2855
	70	15	15	15.68	46.99	48.1	32.3	25.2	46.6	38.1	0.399	0.219	0.273	0.2996
	70	20	10	15.65	46.65	48.0	31.9	26.3	46.3	38.1	0.401	0.221	0.272	0.2934
	70	25	5	15.68	46.86	**48.3**	32.7	25.8	**47.1**	**38.5**	**0.450**	**0.237**	0.275	0.2868

Table 2: **Extensive grid search of data mixing ratios across three axes.** We evaluate models on Language, Visual Understanding, and Visual Generation. The searches confirm that while language requires a dominant token share, visual capabilities peak at highly specific and asymmetrical ratios. The optimal configuration emerges in the "Next" sweep at a 70/25/5 split for Language, Understanding, and Generation respectively.

requires strong understanding and language priors, we derive highly asymmetric data mixing recipes for unified pretraining at scale.

A.4 Architecture Designs in Unified Models

Within the design space of unified multimodal models, maximizing cross-modal synergy while minimizing modality competition remains a central challenge. A critical debate persists regarding the optimal degree of parameter sharing. Some approaches utilize a monolithic transformer for all modalities (Zhou et al., 2025; Wu et al., 2025b; Ma et al., 2025; Chen et al., 2025b; Wu et al., 2024a), whereas others forcibly isolate parameters via modality-specific feed-forward networks (FFNs) (Shi et al., 2024; Lin et al., 2024) or separate attention blocks (Liang et al., 2024; Deng et al., 2025). At the most decoupled design, frameworks can maintain highly separated networks, sharing almost no parameters and passing LLM-derived text features as conditioning signals to an independent diffusion model for generation (Wu et al., 2025a; Chen et al., 2025c; Pan et al., 2025). Our investigation explores how task complexity and architectural choices dictate competition and synergy during unified training. By localizing the exact sources of modality interactions, we confirm that shared attention promotes synergy, while decoupled FFNs prevent competition, providing strong empirical support for MoE architectures (Team, 2026b; Tong et al., 2026a). Furthermore, the design for vision encoders and decoders remains an open question (Han et al., 2026a; Fan et al., 2025; Tong et al., 2026b; Team, 2026b; Liu et al., 2026); to explore this, we systematically investigate three prevalent vision representation designs and evaluate their specific impacts on knowledge flow.

B Data Mixing Ratio Grid Search Details

Table 2 provides the complete results of our extensive grid search over data mixing ratios across 1T token pretraining runs, as discussed in §6.1.

C Extended Discussion, Limitations, and Future Work

Why Generative Synergy is Task-Selective. Our findings in § 3.2.2 and § 3.2 help resolve a common contradiction in multimodal learning: why joint generative pretraining often provides little benefit (or even an alignment penalty) on standard understanding tasks, yet sometimes improves understanding performance in spatial, geometric, and physical reasoning related tasks. Standard benchmarks focus mostly on high-level semantic classification, which only requires sparse and abstract representations. For such tasks, the dense, pixel-level details learned from generation are mostly redundant. In contrast, spatial and structural tasks demand a precise understanding of depth, boundaries, and perspective. The generative objective forces the model to learn these dense visual features, building a richer geometric representation. As shown in our results on low-level property priors, these generation-learned priors can be effectively utilized within understanding streams. Consequently, while this geometric foundation may remain dormant in zero-shot settings, it serves as a highly receptive prior that can be rapidly activated to advance understanding tasks that require dense features.

Optimization Landscapes and Shortcuts in Vision Laziness. As revealed in § 5.3, the stronger the underlying LLM, the more adept it becomes at leveraging its advanced language capabilities to "shortcut" visual problems, a tendency heavily characterized as visual-language prior bias. While relying on such language heuristics is not detrimental for all tasks, it prevents the model from treating vision as an equally important, first-class modality. When the language trunk is already highly optimized before vision is introduced, the joint optimization landscape is heavily dominated by existing language priors. This phenomenon explains why many multimodal LLMs still occasionally act blind or make elementary visual errors: their vision pathways are simply not fully engaged. In this regime, optimization trajectories are dominated by pre-existing representational shortcuts. Early unification is therefore not merely a scheduling choice, but a mechanical necessity required to force the active, unified co-evolution of modalities from scratch.

Internal Mechanics of Parameter Sharing. As analyzed in § 4.2, shared attention and normalization act as a bridge for synergy: shared attention allows modalities to contextualize each other within a joint representational space, and shared normalization aligns their feature scales. Conversely, FFNs act as modality-specific experts that absorb the distinct computational demands of each stream, preventing destructive interference in the shared substrate. Decoupling solely the FFNs perfectly mitigates this competition while leveraging shared attention to foster strong cross-modal synergy, providing direct empirical support for recent Mixture-of-Transformers (MoT) and Mixture-of-Experts (MoE) designs.

Future Discussion. A defining takeaway from our study is the highly efficient, scalable manner in which diverse modalities can co-evolve. Rather than treating generative modeling as an isolated, resource-heavy objective to be aligned post-hoc, we show that it can be powerfully bootstrapped. Because discriminative understanding and language act as strong foundational priors, our optimized recipes suggest that strong generative fidelity can be achieved with only a small fraction of the training budget. By demonstrating that generative capabilities can be unlocked at a low computational cost, we show that native generation can be seamlessly integrated into general-purpose foundation models without the traditional computational penalty. This asymmetric scaling strategy, combined with decoupled parameter allocation and early-fusion co-evolution, provides a concrete, computationally efficient pathway to future models supporting any-to-any I/O.

Looking forward, resolving the delicate balance between synergy and competition becomes even more critical as we transition toward more omnimodal foundation models supporting arbitrary input-output formats (Team, 2026a; Yang et al., 2026a; Tong et al., 2026a). The integration of vision established in this work should naturally extend to continuous video streams, audio, 3D, actions, and even more modalities. By applying our findings on early joint training and parameter decoupling to these richer sensory streams, we can move closer to general-purpose models with highly versatile I/O capabilities. Ultimately, these pretraining principles provide a scalable blueprint for systems with native, high-dimensional I/O, establishing a foundation for a more unified and expressive form of multimodal intelligence. We envision these principles serving as a stepping stone toward scalable multimodal modeling, driving the evolution of multimodal systems from passive multimodal observers into predictive, interactive world models.

Limitations and Future Work. While this work provides a systematic, bottom-up exploration of unified multimodal pretraining, it is subject to several limitations that open exciting avenues for future research. Because our insights are derived from the fundamental aspects of modality interactions, we hypothesize that our core principles (such as asymmetric knowledge flow, the necessity of early unification, and architectural decoupling) will largely hold true for more complex modalities and extreme scales. However, these frontiers remain to be empirically validated.

First, our empirical study is mostly confined to the domain of text and static images, leaving the exploration of dynamic and continuous modalities, such as video and audio, as an open question. Video introduces an additional temporal dimension, which alters data complexity and computational demands. While we expect our static-image findings to serve as a highly transferable baseline, investigating exactly how the principles of knowledge flow and modality synergy apply to video generation and understanding, and whether temporal reasoning requires different architectural decouplings or data mixing curricula, remains a crucial next step for developing predictive world models.

Second, while our proposed pretraining recipes are validated at a substantial scale (e.g., 13.5B parameters trained on 2T tokens), the underlying physics of modality interaction may continue to evolve at the extreme scales of frontier models (e.g., more than 1T parameters). As parameter capacity increases, the threshold at which modalities transition from synergy to competition might shift. More intriguingly, extreme scaling could unlock entirely new forms of cross-modal interactions. For instance, while we currently observe a strict asymmetry where visual generation provides minimal backward transfer to understanding, a massive model might possess enough capacity to utilize generative modeling as an internal world simulator. At such scales, emergent bidirectional knowledge flows could occur, allowing generative processes to actively assist in complex physical reasoning or advanced language tasks. Future work may explore these phenomena at larger scales to determine whether unified pretraining recipes require further recalibration at the frontier.

D Detailed Experimental Setup

D.1 Extended training protocol and tokenization

In this subsection, we detail the complete training parameters and the four visual tokenization configurations evaluated throughout our study:

1. **RAE (default)**: a frozen SigLIP-2 ViT-400m/14 encoder (Zhai et al., 2023; Tschannen et al., 2025) processes 224×224 images into a 16×16 grid of 256 semantic tokens used for both understanding and generation, with generation performing flow matching in the SigLIP latent space and decoded by a Representation Autoencoder (Zheng et al., 2026).
2. **Raw Pixels**: the encoder/decoder is removed; a 224×224 image is patchified by a single 14×14 convolution into a 16×16 grid of 256 patch tokens, projected by a lightweight MLP to the transformer hidden dimension (with bilinear resampling to align the token count when needed), and the same tokens are consumed for understanding and denoised for generation.
3. **CLIP + VAE**: SigLIP-2 is used for understanding at 224×224 (256 tokens), while generation operates at 256×256 in the Stable Diffusion 3 VAE latent space ($8\times$ downsampling, 16 channels), denoising a 32×32 latent grid.
4. **AR (UniTok)**: to evaluate whether our findings generalize beyond diffusion-based image modeling, we implement an autoregressive configuration using discrete visual codes from UniTok (Jiao et al., 2025). Images are represented by residual-quantized codes, which are predicted autoregressively across spatial positions. At each position, a dedicated causal depth head predicts the codebook factors sequentially under a cross-entropy objective.

Training is optimized using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay of 0.1, and gradient clipping at 1.0, following a cosine decay schedule with a linear warm-up over the first 8000 steps. To balance cross-modal learning, the continuous diffusion flow-matching loss is up-weighted by a factor of 3.0 relative to the discrete text cross-entropy loss. Training timesteps $t \in [0, 1]$ are sampled from a logit-normal distribution. Models are trained with a context length of 4096 tokens in bf16 precision under FSDP-2. Our controlled experiments span training budgets ranging from 100B to 2T tokens.

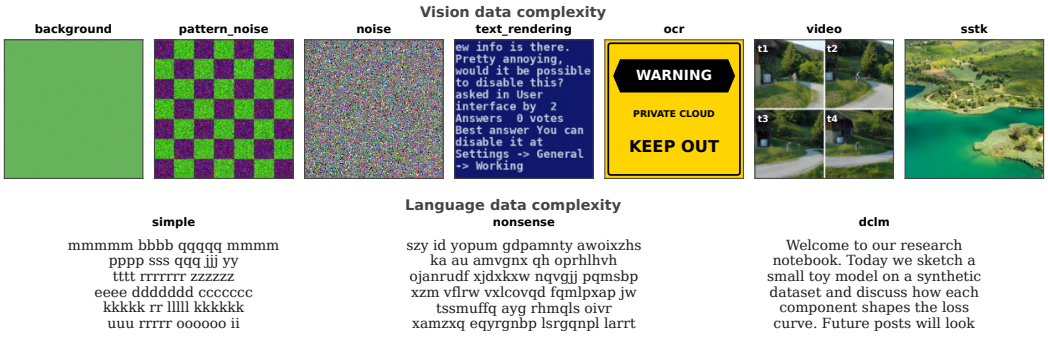


Figure 11: **Overview of data complexity progressions.** Examples of visual (top) and language (bottom) data used to evaluate the impact of task complexity on modality interactions, ranging from simple synthetic patterns to complex real-world distributions.

Following the visual question answering (VQA) evaluation protocol established by Cambrian-1 (Tong et al., 2024a), Web-SSL (Fan et al., 2025), LSBS (Han et al., 2026b), and Beyond Language Modeling (Tong et al., 2026a), we perform one epoch of supervised fine-tuning on the Cambrian-7M instruction dataset after pretraining to obtain visual-understanding scores. Fine-tuning uses AdamW with a peak learning rate of 1×10^{-5}, cosine decay, weight decay of 0.1, and an effective global batch size of 128 sequences. We use the full Cambrian-7M curation as the supervised fine-tuning corpus, mixing language-only and vision-language paired instructions in their original proportions.

D.2 Detailed Evaluation Protocol

Language evaluation. We measure language capability along two complementary axes: few-shot downstream accuracy and validation perplexity. The downstream score is the unweighted average over **11 standard benchmarks**, comprising nine multiple-choice reasoning and commonsense tasks scored by accuracy, including ARC-Easy and ARC-Challenge (Clark et al., 2018), BoolQ (Clark et al., 2019), CoQA (Reddy et al., 2019), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021), and two open-ended QA tasks scored by exact match, NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). We also report validation average perplexity on DCLM (in-distribution) and C4 (out-of-distribution).

VQA evaluation. We comprehensively evaluate visual capabilities across understanding and generation. For visual understanding, we categorize our 16 benchmarks into four core evaluation axes:

- **General:** Focuses on basic visual perception and its alignment with commonsense, avoiding complex inferential tasks. Evaluated on GQA (Hudson & Manning, 2019), MME (Fu et al., 2025), MMBench (Liu et al., 2024), and SEED (Ge et al., 2023).
- **Knowledge:** Probes the integration of visual cues with broad world knowledge, demanding multi-step reasoning for scientific and mathematical problem-solving. Covers ScienceQA (Lu et al., 2022b), MMMU (Yue et al., 2024), AI2D (Hiippala et al., 2021), and MathVista (Lu et al., 2023).
- **OCR & Chart VQA:** Assesses high-resolution, fine-grained perception necessary for reading dense textual elements and decoding structured charts. Comprises TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and OCRBench (Liu et al., 2023b).
- **Vision-Centric:** Tests native visual faculties such as spatial understanding, 3D perception, object counting, and abstract visual logic. Benchmarks include RealWorldQA (xAI, 2024), MMVP (Tong et al., 2024b), CV-Bench (derived from Cambrian-1 (Tong et al., 2024a), consisting of COCO (Lin et al., 2014), ADE (Zhou et al., 2019), and Omni3D (Brazil et al., 2023)).

Visual generation evaluation. We measure compositional text-to-image capability using GenEval (Ghosh et al., 2023), and evaluate dense-prompt generation via DPG-Bench (Hu et al., 2024). For text-image alignment, we report CLIP similarity scores (Hessel et al., 2021) across three prompt-length buckets—short ($<$10 words), medium (10–30 words), and long (30–50 words), each comprising a few hundred prompts. Finally, as a proxy for overall generation quality, we report the held-out diffusion loss measured on a validation set containing 1000 samples.

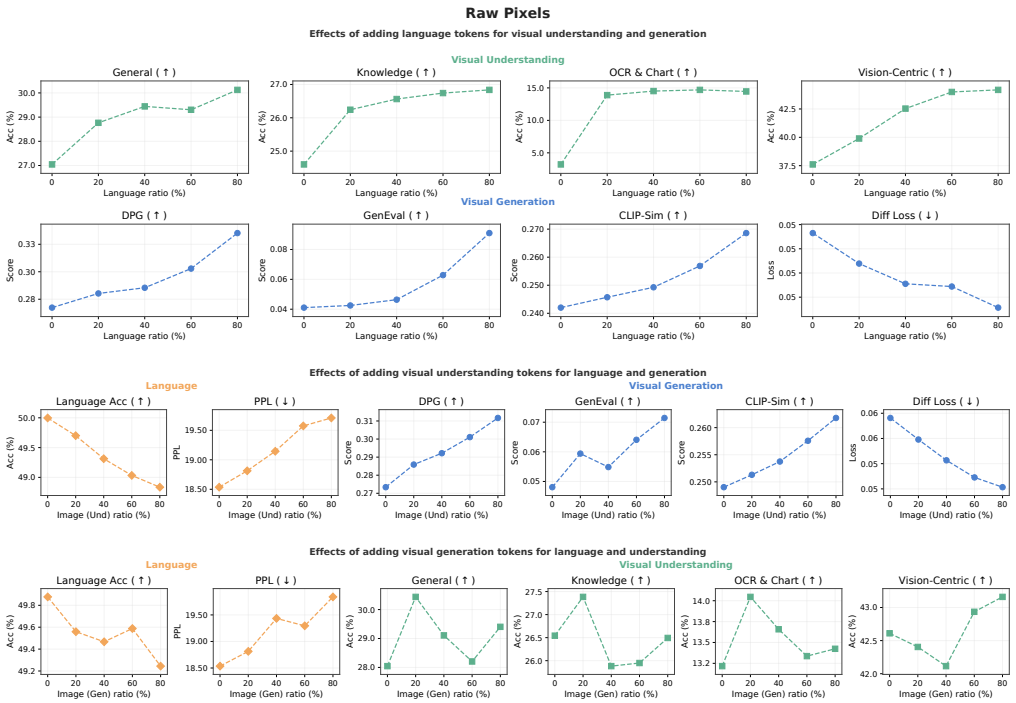


Figure 12: **Modality transfer results using Raw Pixels.** We replicate the real-world knowledge flow transfer experiments using raw pixels. The results mirror the main text: language universally boosts vision (top), understanding significantly benefits generation (middle), and generation yields a highly stable, neutral effect on language and visual understanding tasks (bottom).

E Modality Transfer on Real-world Data Additional Results

In §3.1 of the main text, we analyze the modality knowledge flow using our default RAE visual tokenizer/encoder. To verify that our findings on knowledge flow asymmetry and neutrality are applicable to different visual tokenizer designs, we replicate the real-world modality transfer experiments under three alternative vision designs: (1) **Raw Pixels**, (2) **CLIP + VAE**, and (3) **AR (UniTok)**. For these supplementary transfer evaluations, we primarily focus on conditional visual generation.

As illustrated in Figure 12 (Raw Pixels), Figure 13 (CLIP + VAE), and Figure 14 (AR modeling), the empirical results exhibit highly consistent behaviors across all three tokenization designs. Across all designs, we consistently observe:

- **Language as a universal booster:** Across alternative setups, increasing the language data ratio from 0% to 80% monotonically and significantly improves all visual understanding metrics as well as generative quality.
- **Understanding as a prior for generation:** Scaling the proportion of visual understanding tokens consistently lowers the diffusion loss and enhances text-to-image alignment scores under three settings, confirming the effective transfer of visual understanding knowledge to generative processes.
- **Neutral effect of visual generation:** Increasing the visual generation compute budget yields a largely neutral, highly stable effect on other capabilities. Rather than triggering any systematic degradation, both language and visual understanding benchmarks exhibit only minor, non-directional fluctuations across the entire range of generation ratios.

These supplementary evaluations confirm that modality knowledge flows are more tied to the learning objectives themselves rather than specific tokenization designs.

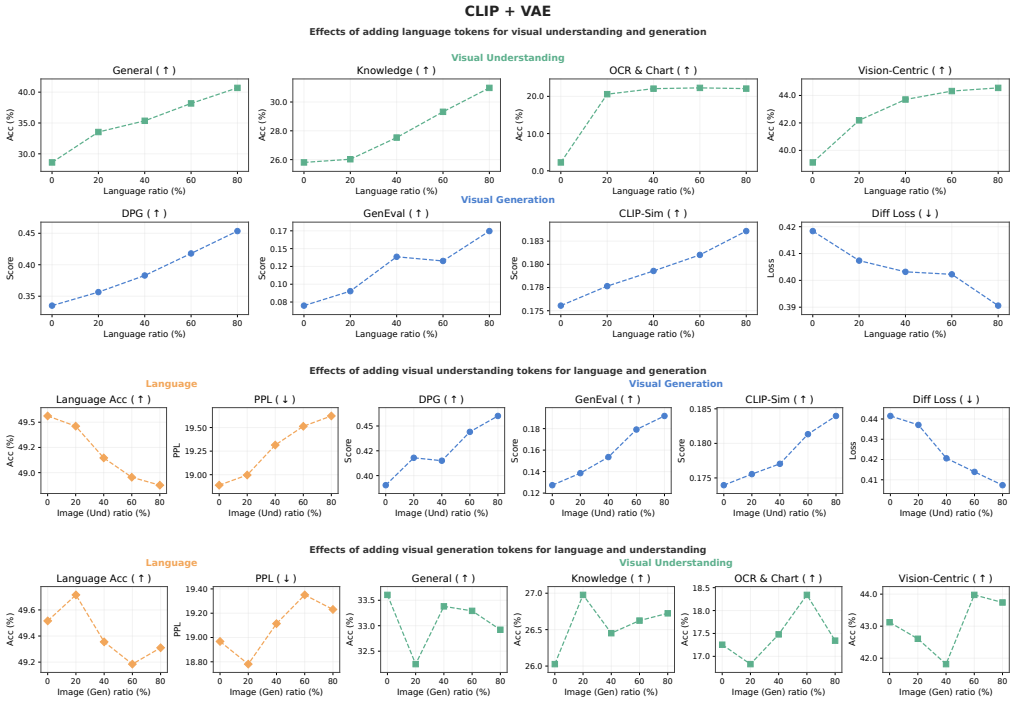


Figure 13: **Modality transfer results using CLIP + VAE.** We replicate the real-world modality transfer experiments utilizing a SigLIP-2 encoder for visual understanding and an SD3 VAE for generation. The transfer dynamics remain highly consistent, showing that language acts as a universal booster, understanding acts as a strong prior for generation, and generation has a stable, neutral, or slightly fluctuating impact on other abilities.

F CLEVR Generation and Evaluation Details

F.1 Generation

Synthetic dataset environment. We build our synthetic dataset on top of the CLEVR rendering pipeline. To increase the complexity of visual features, we extend the standard CLEVR asset library by expanding the color and shape vocabularies:

- **Expanded Colors:** Standard colors plus additional custom hues, including `red`, `blue`, `green`, `gray`, `brown`, `purple`, `cyan`, and `yellow`.
- **Expanded Shapes:** Standard primitives (`cube`, `sphere`, `cylinder`) along with newly authored shapes: `cross`, `ellipse`, `pentagon`, `rectangle`, and `triangle`.

Blender renders each scene based on a programmatically sampled scene graph that defines the color, shape, size, and 3D position of every object. Each scene contains between 1 and 7 objects. To ensure clean visual rendering, we programmatically enforce distance constraints during placement to prevent physical intersections or visual overlap. In total, we generate roughly one million scenes for the training mixture and reserve a held-out set as the evaluation split.

Dense and sparse captions. For each rendered image, we generate a dense caption (prompted to detail every object's shape, size, color, and position) and a sparse caption (prompted to briefly describe the image) using Qwen3-VL-8B-Instruct (temperature 0.2, top-p 0.8, a maximum of 256 tokens). We mix dense and sparse captions together during training. Each caption serves both visual understanding (image $\rightarrow$ text) and visual generation (text $\rightarrow$ image).

Five target concepts. To analyze knowledge flow, we focus our systematic ablation and evaluation on five specific conceptual axes:

1. **Color:** Discriminating or rendering the color of target objects.

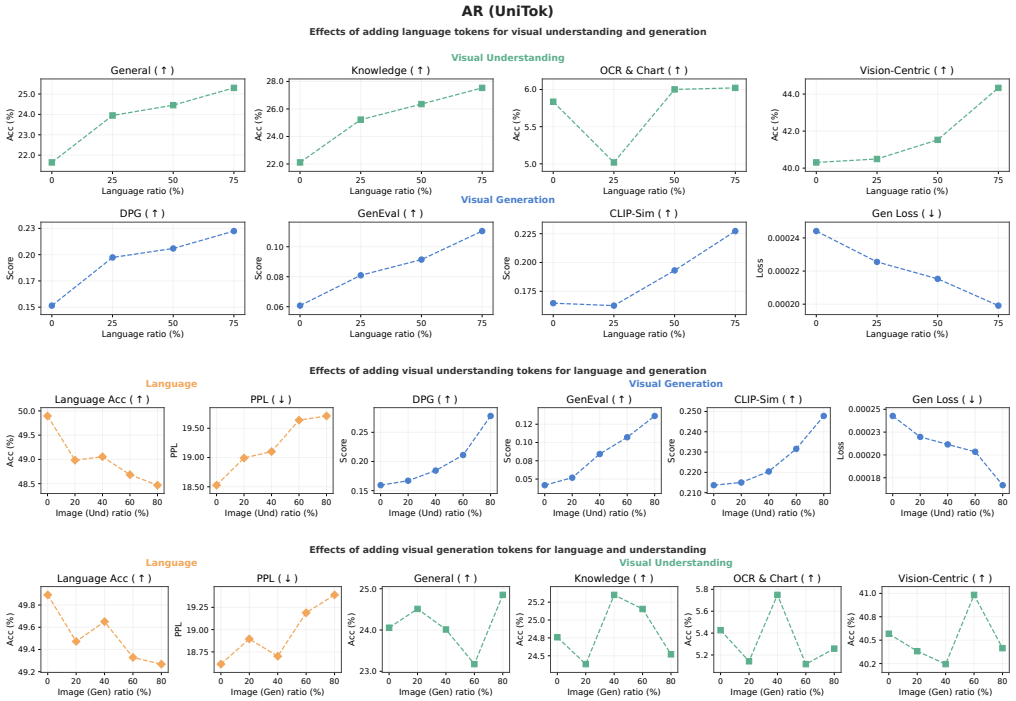


Figure 14: **Modality transfer results with AR modeling.** We replicate the real-world modality transfer experiments utilizing discrete visual tokenization and autoregressive next-token prediction. Consistent with our continuous diffusion findings, the results demonstrate that language acts as a universal booster (top), visual understanding serves as a strong prior to drive generation (middle), and visual generation exerts a highly stable, neutral effect on other capabilities (bottom).

2. **Shape:** Recognizing or generating the precise geometry of objects.
3. **Relation:** Comprehending spatial directional offsets between objects. We compute the ground-truth directions analytically using the 3D coordinate metadata in the scene graph to support both cardinal and diagonal relations (e.g., `left`, `right`, `front-left`, `behind-right`).
4. **Size:** Categorizing absolute object scales (`large`, `medium`, `small`) based on the rendered dimension, where questions query either the absolute scale of a single object or relative comparisons between multiple objects.
5. **Count:** Quantifying the total number of objects in the scene.

VQA pair construction. We derive VQA pairs directly from the scene graph metadata, ensuring noise-free ground truth by construction. We format these pairs as open-ended, short-answer questions that revolve around the five target concepts. Specifically, we generate questions across several structural types:

- **Attribute identification** queries the color, shape, or scale of a target object (e.g., "What is the shape of the blue object?"). We enforce a uniqueness filter on the referent attributes to ensure the target object is uniquely identifiable, discarding the question if ambiguity exists.
- **Counting and size comparison** queries the total number of objects in the scene (e.g., "How many objects are in the image?"), the count of objects matching specific descriptors (e.g., "How many green cubes are there?"), absolute size categorization, or relative size comparisons between objects (e.g., "Is the sphere larger than the cylinder?").
- **Spatial reasoning** covers single-hop and multi-hop spatial relationships. For single-hop queries, we directly ask for the directional orientation of one object relative to another (e.g., "Where is the sphere relative to the cube?"), where camera-relative directions are analytically computed using 3D coordinate offsets to produce diagonal outputs such as `front-left` or

`behind-right`. For multi-hop queries, we precompute directional neighbors to formulate sequential paths (e.g., "What color is the object behind the object to the left of the cube?").

The corresponding answers are formatted as single-word or short-phrase text generation targets (e.g., `red`, `cylinder`, `three`, or `front-left`). To prevent training bias, we balance the question generation process so that the total number of VQA pairs for each of the five target concepts is roughly equal. Each instance derived from a rendered scene consists of one image, two captions (dense and sparse), and one to three VQA questions.

F.2 Evaluation

We evaluate concept transfer using a controlled leave-one-concept-out study across two symmetric directions. For every held-out concept c, we use 100 evaluation samples.

Understanding → Generation. For each held-out concept value c (e.g., color `yellow`, shape `sphere`, relation `front-left`, count `4+`, size `large`), we surgically remove any scene whose scene-graph metadata contains the concept c from the generation training data. Consequently, the model never observes any training image containing the ablated target (e.g., no yellow objects), while we leave the understanding data (captions used as image → text inputs, and VQA pairs) completely untouched.

We then ask whether the model can still produce c when prompted, by feeding it ablate prompts that explicitly request the held-out concept. For $c =$ `yellow`, for example, every ablate prompt mentions a yellow object, sampled from templates such as "*A yellow triangle is on the front-right side of the image*" or "*A yellow rectangle is to the right of a cyan ellipse*"; for $c =$ `sphere`, each prompt places a sphere in the scene; for $c =$ `front-left`, each prompt requests a front-left arrangement; for $c =$ `4+`, each prompt requests a scene containing at least four objects; and for $c =$ `large`, each prompt requests a large object. The neutral control set draws from the same templates but uses values that remain present in the training mixture.

We score the resulting images using a VLM-judge protocol. Specifically, Qwen3-VL-8B-Instruct directly processes the generated image alongside the original generative prompt, answering a binary yes/no question targeting the held-out axis (e.g., "does the image depict the correct COLOR / SHAPE / RELATION / SIZE / COUNT?"). We report per-axis accuracy on the ablate set restricted to the axis that we hold out, making the metric directly sensitive to whether the concept successfully appears in the generated pixels.

Generation → Understanding. The symmetric study removes c from the understanding stream (both captions and VQA pairs that reference c) while leaving the generation stream untouched. We then ask whether the model can still answer questions about c at evaluation time. Concretely, we draw the 100 ablation VQA examples from the held-out scenes whose ground-truth answer is c, meaning the model only scores a point when it generates c itself.

For $c =$ `yellow`, this setup includes questions such as "*What is the color of the object?*" on a single yellow-object scene, "*What is the color of the triangle?*" on a multi-object scene whose unique triangle is yellow, or "*What is the color of the object to the left of the gray cube?*" . We apply the same construction to the other concepts: for $c =$ `sphere`, we sample shape questions whose answer is `sphere`; for $c =$ `front-left`, we sample spatial-relation questions whose ground-truth direction is `front-left`; for $c =$ `4+`, we sample counting questions whose ground-truth count is `four`; and for $c =$ `large`, we sample size questions whose ground-truth label is `large`. We report the average VQA accuracy over these 100 questions for each concept.

G CLEVR Additional Results

In this section, we provide supplementary quantitative and qualitative results on the synthetic CLEVR dataset to support the analyses and findings presented in §3.2.

First, to offer a more granular look at the concept prior study, we plot the detailed training trajectories across the first 1000 steps of optimization process. Figure 15 displays the step-by-step diffusion loss curves when recovering held-out low-level concepts (color and shape) for the generative task under the presence of understanding priors, serving as the supplementary result for the top row of Figure 6.

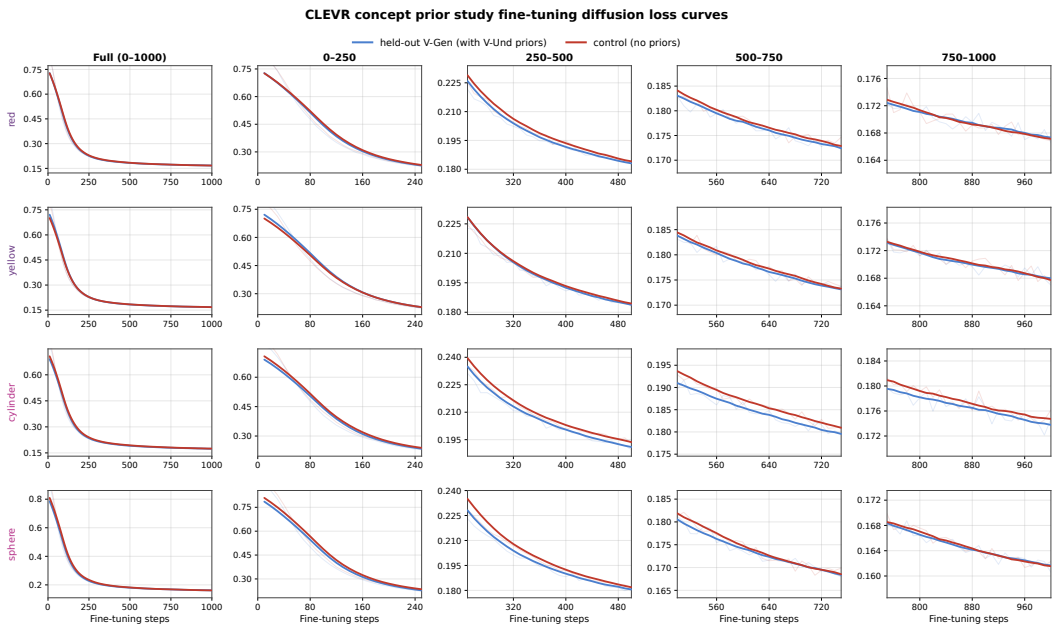


Figure 15: **Fine-tuning diffusion loss curves for concept recovery.** We show the training trajectories of the diffusion loss during fine-tuning on the generative task for four low-level concepts: red, yellow, cylinder, and sphere. The curves compare the model with visual understanding priors (blue lines) against the control group with no prior exposure (red lines) across different step intervals. The trajectories show that prior exposure via visual understanding provides no observable benefit for subsequent color generation (overlapping closely), but yields a marginal, minor acceleration for shape generation.

Symmetrically, Figure 16 shows the text cross-entropy loss curves when utilizing generative priors to recover these same concepts for the understanding (VQA) task, serving as the extended data for the bottom row of Figure 6.

Second, to complement the quantitative zero-shot transfer rates shown in Figure 5 and discussed in §3.2.2, we provide qualitative visual examples in Figure 17. These samples showcase the model's generation capabilities under different ablation settings, visually illustrating where structural knowledge transfer succeeds (such as spatial relations, size, and object count) and where low-level semantic transfer fails (such as color and shape).

H Modality Synergy Across Multiple Vision Encoder Designs

To evaluate whether the observed cross-modal synergies and competition dynamics are tied to a specific visual representation, we systematically investigate how different vision encoder and tokenization choices affect these interactions.

Setting. Following the `split_ffn` training protocol from §4.2, we pretrain models on a 100B token budget (50B language / 50B visual) using our synergy-maximizing probes: pairing pure background images with language to measure the change in language perplexity (Δ PPL), and pairing simple repetitive language with images to measure the relative reduction in visual generation loss (relative Δ diffusion loss %). We benchmark these probes across the four visual tokenization designs introduced in §2:

1. **RAE (default)**: Continuous diffusion in frozen SigLIP-2 ViT-400m/14 semantic latent space, decoded via Representation Autoencoders.
2. **Raw Pixels**: Direct patchification of 224×224 images into 256 patch tokens without semantic pretraining, jointly consumed for understanding and flow-matching denoising.
3. **CLIP + VAE**: Segregated representation spaces using SigLIP-2 for understanding and SD3 VAE latents (32×32) for generative denoising.

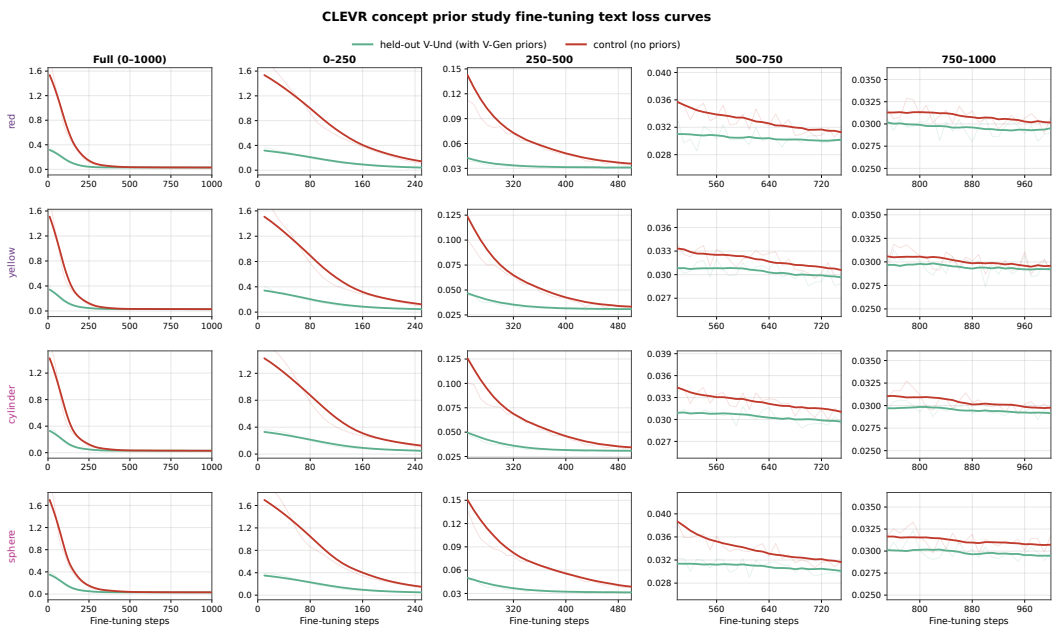


Figure 16: **Fine-tuning text loss curves for concept recovery.** We show the cross-entropy training loss trajectories during fine-tuning on the understanding (VQA) task for four low-level concepts: red, yellow, cylinder, and sphere. The curves compare the model with visual generation priors (green lines) against the control group with no prior exposure (red lines). These results demonstrate that generative training establishes fine-grained priors in low-level attributes (e.g., pixel-level color and texture distributions) for visual understanding. While not immediately accessible zero-shot, these latent representations serve as strong priors that accelerate learning when mapping visual features to visual understanding.

4. **AR (UniTok)**: Discrete sequence modeling predicting residual-quantized visual tokens autoregressively under cross-entropy.

Results and Discussion. As shown in Figure 18, cross-modal synergy reliably manifests across all four architectural paradigms:

- **Raw Pixels:** Even in the complete absence of semantic pretraining or visual abstractions, pairing raw pixels with pure backgrounds yields the largest language perplexity improvement (Δ PPL -0.266), surpassing the default RAE setup (-0.211). Similarly, generation diffusion loss drops by -2.19% (conditional) and -2.82% (unconditional) when trained with simplified language.
- **Decoupled CLIP + VAE:** When understanding and generation reside in entirely separate latent spaces, cross-modal synergy still persists. While the text perplexity gain is more modest (Δ PPL -0.024), visual generation receives a substantial boost (relative diffusion loss drops by -2.24% and -2.83%).
- **Autoregressive Discrete Modeling:** The UniTok autoregressive setup exhibits consistent synergy trends, verifying that modality cooperation is not an artifact of continuous flow matching or diffusion objectives.

Combined with our real-world modality transfer experiments (Appendix E, Figures 12–14), these findings clarify a central question in unified multimodal modeling: *is a unified, shared visual representation space strictly required for cross-modal synergy?*

Our results demonstrate that a pre-aligned or unified visual latent space is **not an absolute prerequisite** for cross-modal synergy or asymmetric knowledge flow; rather, modality synergy is primarily governed by training task objectives, data complexity, and parameter-sharing strategies within Transformer blocks. Nonetheless, adopting a shared semantic latent space (like RAE) remains highly advantageous in practice, as it minimizes representational friction, reduces architectural complexity, and maximizes training compute efficiency.

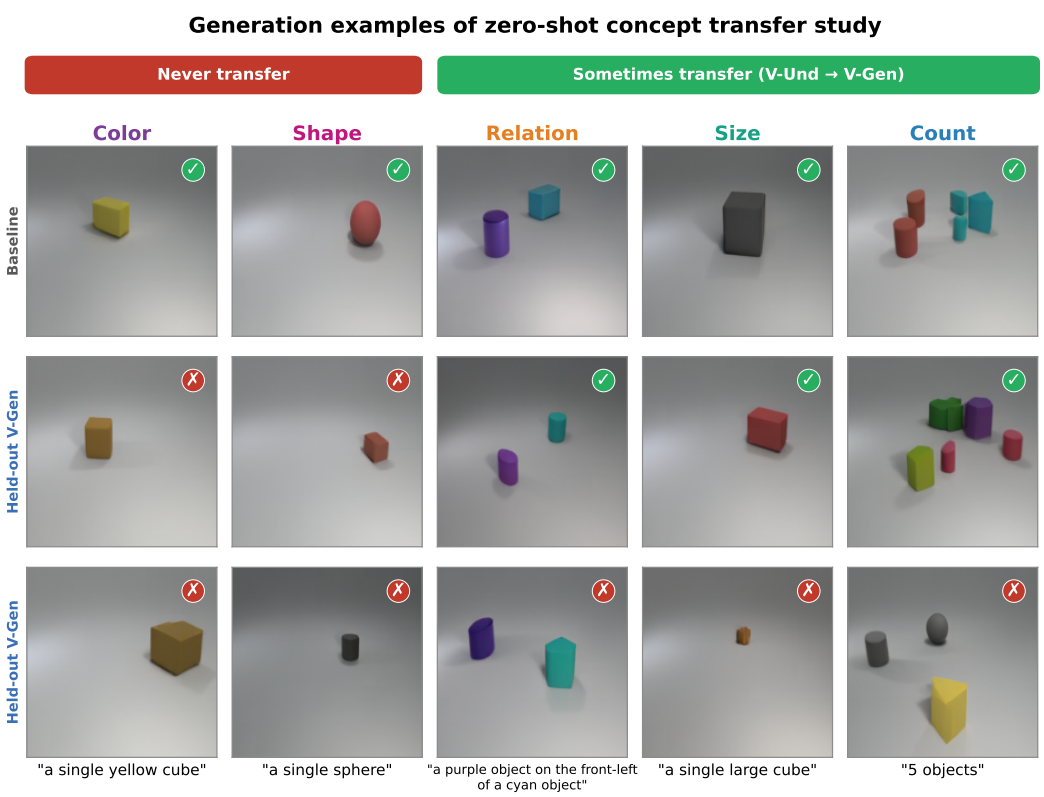


Figure 17: **Qualitative examples of zero-shot concept transfer.** We present synthetic CLEVR scenes generated under different pretraining configurations. The baseline model (top row) represents joint training on all modalities and concepts. For models where specific concepts are held out from the generation stream but preserved in the understanding stream (middle and bottom rows), zero-shot generation exhibits partial transfer, sometimes successfully rendering higher-level structural concepts such as spatial relations, size, and object count (green checkmarks), but consistently failing to generate low-level attributes like color and shape (red crosses), showcasing the asymmetric, concept-dependent nature of cross-modal transfer.

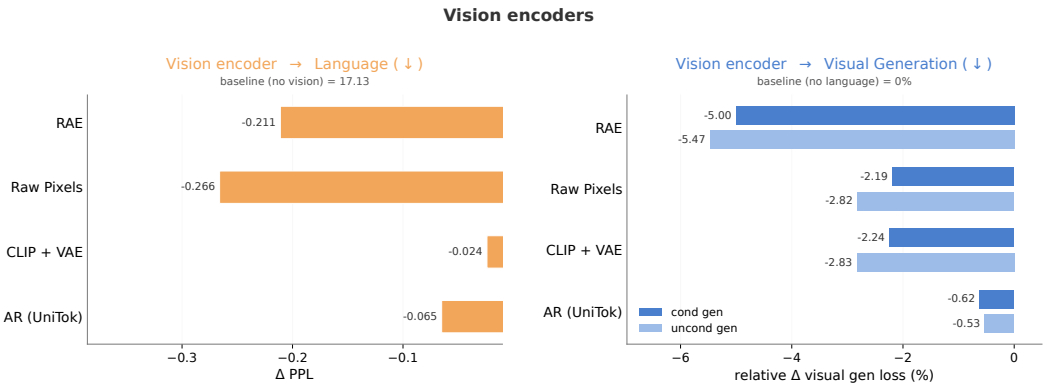


Figure 18: **Impact of vision encoder designs on modality synergy. Left:** Language perplexity change (Δ PPL) when paired with pure background visual inputs across different encoder configurations. **Right:** Relative reduction in generation diffusion loss (%) for conditional and unconditional generation when paired with simplified language. Modality synergy consistently emerges across all four tokenization paradigms.

I Additional Text-to-Image Generation Results

We qualitatively compare text-to-image generation between our early unification model (Balanced Recipe) and the late alignment baseline (Late Fusion). Both are 13.5B MoE models trained under the matched compute budget described in § 6.2. Figure 19 showcases their generated images across diverse prompts.

Qualitative analysis of generative samples. A direct comparison of the generated samples shows noticeable differences in how both models handle prompt instructions. For multi-attribute binding prompts, such as *"Italian noodles in three colors..."* or *"Duck with green head sitting on bright green grass,"* the early-fusion model tends to maintain clearer boundaries between distinct elements. In the noodle example, it generates separate red, white, and green segments, whereas the late-fusion model tends to blend these colors together into a generic pasta dish. Similar trends appear in geometric and symbolic representations like *"Bitcoin crypto currency sign element made of clouds,"* where the early-fusion model outlines a recognizable shape while the late-fusion model outputs standard unstructured clouds. Additionally, for prompts demanding specific contextual elements such as reflections or hanging objects, the early-fusion model attempts to render these details more consistently, whereas the late-fusion model sometimes omits them or renders them as highly ambiguous shapes.

Vision laziness in visual generation. These qualitative variations are consistent with the vision laziness phenomenon discussed in § 5.3. When the language pathway is allowed to harden during an extended pure-language pretraining phase, the model may fail to fully optimize its visual pathways, potentially resulting in the simpler geometries and less defined textures observed in the late-fusion outputs. This possible under-optimization of the visual components is highly aligned with the quieter activations shown in Figure 10. Consequently, the late-fusion model appears to rely primarily on coarse language priors, rendering a standard subject while neglecting specific descriptive adjectives like *"green head"* or *"nicely reflected."* While qualitative samples cannot definitively prove the underlying optimization dynamics, the observed degradation in multi-attribute binding and structural details in late-fusion aligns well with the quantitative measurements of reduced visual pathway activity.

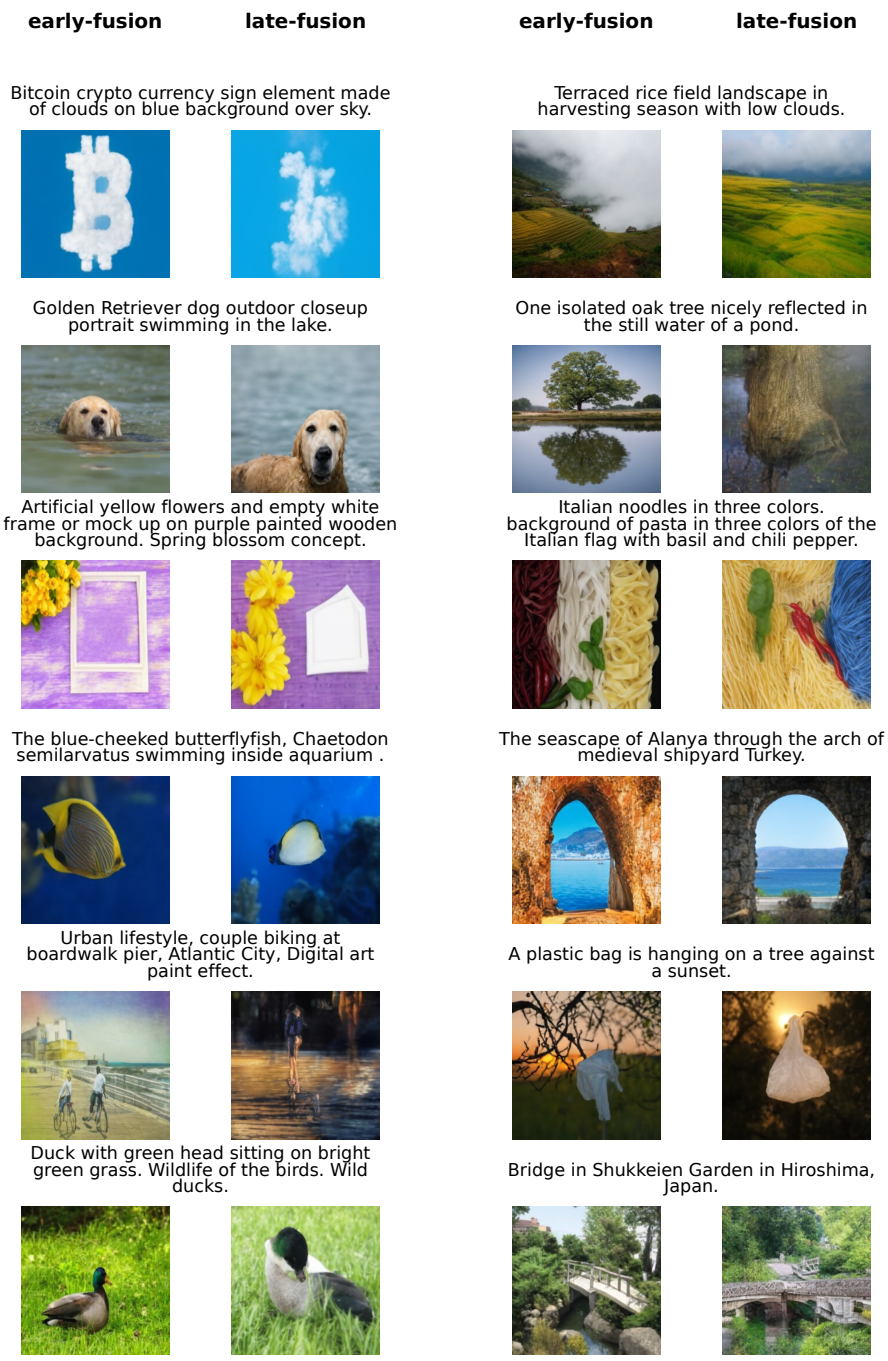


Figure 19: **Qualitative comparisons of text-to-image generation between early-fusion and late-fusion.** The early-fusion model generally exhibits closer adherence to multi-attribute constraints and sharper geometric structures, whereas the late-fusion baseline tends to produce simplified or abstract shapes.