CONFLUX: ROLE SEPARATION FOR UNIFIED MULTIMODAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models with a shared backbone face costly backbone iterations for image generation and conflicting visibility requirements for prediction and conditioning. We introduce CONFLUX, which separates semantic conditioning from rendering and observed content from prediction queries. The backbone computes image conditions once per target image and conditioning branch; a diffusion transformer reuses them for iterative rendering. We adopt content–query factorization across modalities with a common representation interface: queries encode leakage-free prediction conditions, while content states preserve complete, clean image context for subsequent predictions. With a fully trainable 0.6B backbone and matched data and update budgets, CONFLUX improves MMLU by 7.56 points and ImageNet classification by 3.53 points over the strongest respective baselines in the main comparison. Image generation achieves $4.0\times$ the throughput of the fastest baseline under common settings, with comparable FID (3.67 versus 3.09) at separately selected generation configurations. Scaling only the flow head improves FID to 3.19 while largely preserving language and visual understanding performance and retaining a throughput advantage.

1 INTRODUCTION

Unified multimodal models with a shared backbone aim to integrate knowledge and contextual reasoning across language, visual understanding, and image generation. The success of next-token prediction as a general-purpose pretraining objective for large language models motivates a broader question: how can we develop a similarly coherent recipe for training unified multimodal models from scratch? This extension introduces distinct representational and computational demands. A shared backbone must support both semantic modeling and image rendering, while accommodating each modality as both observed context and a prediction target. We investigate this question through architectural choices that balance multimodal modeling capability with computational cost and provide each prediction with a modality-appropriate representation of the available context.

For image generation, existing jointly trained unified multimodal models repeatedly invoke their backbone on intermediate image states: across denoising iterations (Zhou et al., 2025; Xie et al., 2025), image tokens (Sun et al., 2024; Kou et al., 2025), or reveal rounds (Yang et al., 2024), corresponding to T, N, or K backbone iterations in Figure 1a. Processing evolving image states can couple semantic conditioning with rendering, and incurs a generation cost that scales with the product of the iteration count and the backbone size, which is prohibitive if we try to scale up the backbone size to hundreds of billions.

The same designs also constrain how images serve as context during joint training. In an idealized teacher-forced pass over a mixed-modality sequence $[P, I, S]$, with text segments P and S, image I would serve as both the target of $p(I \mid P)$ and context for $p(S \mid P, I)$. These two roles, however, favor different attention structures from our ablations. In existing designs illustrated in Figure 1b, however, reusing noisy, causally restricted, or partially visible image states ties the context available to subsequent predictions to the image generation procedure.

To address both challenges, we introduce CONFLUX, a jointly trained unified multimodal model with two complementary role separations. Our architecture borrows ideas from DreamLLM (Dong et al., 2024) with critical new designs and simplifications.

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

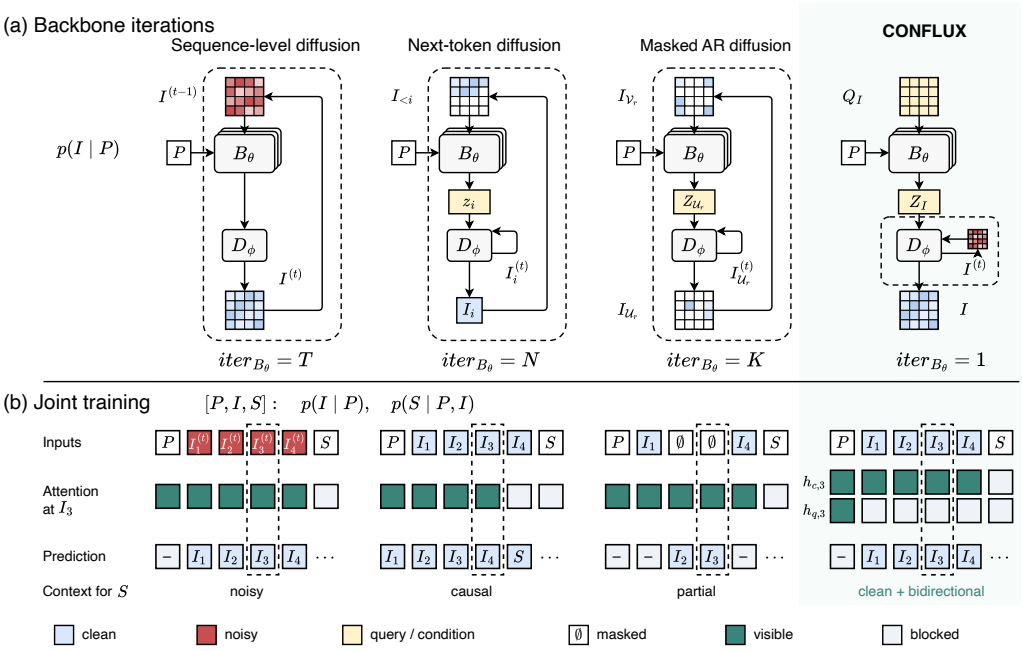


Figure 1: **Image generation and context visibility in unified multimodal models.** (a) The backbone repeats over T denoising iterations, N image tokens, or K reveal rounds in the first three designs. CONFLUX invokes the backbone once and iterates only the diffusion head. Counts are per image and conditioning branch. (b) In an idealized teacher-forced pass over $[P, I, S]$, the three baselines would supply noisy, causal, or partial image context for S. CONFLUX separates complete, clean, bidirectional content from queries that exclude the target image. Dashed boxes highlight I_3.

The first separates *where* computation is performed. The shared backbone is responsible for semantic and contextual modeling, while a lightweight flow transformer performs iterative image synthesis. The backbone computes all target-image conditions in a single forward pass, after which the flow head reuses them throughout generation; noisy latents and flow time remain entirely outside the backbone. This keeps iterative rendering from dominating backbone computation and allows rendering capacity to scale independently of semantic-model capacity. DreamLLM (Dong et al., 2024) adopts a related backbone–renderer decomposition, but uses the multimodal model primarily to condition a fixed pretrained Stable Diffusion U-Net. In contrast, our backbone and compact DiT-based flow head are jointly trained under a unified objective, so semantic representations and rendering co-adapt while remaining computationally separated. Their conditioning interfaces are also different. DreamLLM introduces a fixed bank of 64 independently learned dream queries. CONFLUX instead assigns a condition state to each token, initialized with the same modality embedding.

The second separation determines *how* semantic information is represented inside the shared backbone. Inspired by XLNet's two-stream attention (Yang et al., 2019), we distinguish observed content states from prediction queries for both text and images. Content states provide keys and values and represent information that may serve as context for later predictions, whereas prediction queries read only the content visible before their target and never attend to one another. For text, each content state encodes the observed token and its causal context, while a separate query predicts that token without accessing its content state. For images, content states encode the complete clean image with bidirectional intra-image attention, while parallel queries access only preceding context and produce the semantic conditions consumed by the flow head. Target-specific visibility masks therefore allow the same image to be hidden from its own prediction queries yet exposed in full as clean context to subsequent predictions.

DreamLLM's dream queries provide a related image-generation interface, but remain generation-specific: a fixed bank of 64 learned queries is appended to an otherwise causal backbone, while observed images are represented separately by a frozen CLIP encoder and text uses ordinary causal

states. Our design instead promotes content–query factorization to a modality-general backbone abstraction. The same mechanism governs text prediction, image prediction, and the use of observed multimodal content as context. Together, the two separations yield a unified architecture in which the backbone performs semantic reasoning once, prediction states remain leakage-free, observed images remain fully available as clean context, and iterative rendering is delegated to a small specialized head.

Under matched data and update budgets with a fully trainable 0.6B backbone and 0.16B flow head, CONFLUX improves MMLU by 7.56 points and ImageNet classification by 3.53 points over the strongest respective baselines. It achieves $4.0\times$ the image generation throughput of the fastest baseline under a common sampling configuration, with comparable FID (3.67 versus 3.09). Scaling only the flow head further improves FID to 3.19 while leaving language and visual understanding performance largely unchanged and retaining a throughput advantage over the baselines.

Contributions.

- We introduce CONFLUX, a unified multimodal architecture built around two complementary separations: *semantic conditioning from rendering* and *observed content from prediction queries*. Together, they move iterative image synthesis into a lightweight flow head while letting the shared backbone produce leakage-free text and image prediction states and simultaneously preserve complete clean image context for subsequent predictions in a single teacher-forced pass.
- Under matched data and update budgets, CONFLUX improves pure-text and visual understanding over unified baselines while achieving $4.0\times$ the image-generation throughput of the fastest baseline at comparable FID, demonstrating that stronger multimodal understanding need not come at the cost of efficient image synthesis.
- We show that rendering capacity can be scaled independently of backbone capacity: enlarging the flow head substantially improves image-generation quality while largely preserving language and visual understanding performance and retaining a clear generation-throughput advantage.

2 Related Work

Unified multimodal modeling. Unified multimodal models extend language modeling to visual understanding and image generation within a shared backbone. Chameleon and Emu3 model text and discrete visual tokens with a common next-token objective (Chameleon Team, 2024; Wang et al., 2024). Continuous-latent approaches combine language prediction with image diffusion: Transfusion and Show-o2 jointly model text and image sequences (Zhou et al., 2025; Xie et al., 2025), while LatentLM, Orthus, and MMAR explore autoregressive formulations with diffusion-based image prediction (Sun et al., 2024; Kou et al., 2025; Yang et al., 2024). Our focus is the representation interface within the shared backbone: separating observed context from prediction conditions across modalities, while assigning iterative image rendering to a flow head.

Query interfaces for image generation. DreamLLM (Dong et al., 2024), MetaQueries (Pan et al., 2025), and ScaleRAE (Tong et al., 2026) use learned queries to extract image-generation conditions from a language backbone for a diffusion decoder. Their training strategies freeze different components: DreamLLM keeps its visual encoder and Stable Diffusion decoder frozen, unfreezing the LLM after alignment; MetaQueries freezes the multimodal backbone in its main configuration; ScaleRAE freezes its representation encoder but jointly trains the LLM, queries, and diffusion transformer. These interfaces distinguish observed image content from generation queries while retaining next-token text prediction. For example, DreamLLM uses visual embeddings for observed images and separate dream queries for synthesis; MetaQueries processes generation queries with causal self-attention, allowing earlier queries to supply keys and values to later queries. CONFLUX adopts XLNet's content–query factorization (Yang et al., 2019) across both text and images, with shared backbone parameters. Queries at target positions read keys and values exclusively from observed-content states, without query-to-query attention; interactions among noisy image latents are modeled in the flow head. This yields a common interface for both modalities: queries encode

prediction conditions without accessing their targets, while content states remain available as context for subsequent predictions.

3 PRELIMINARIES

3.1 UNIFIED MULTIMODAL MODELING AND RECTIFIED FLOW

A shared backbone B_θ processes mixed sequences $x = (x_1, \ldots, x_n)$ of text tokens and image latents. A frozen autoencoder maps each image to a grid $I = (I_1, \ldots, I_N) \in \mathbb{R}^{N \times d}$. Given preceding multimodal context c, a categorical head predicts text autoregressively, while a flow head D_ϕ models $p(I \mid c)$.

Rectified flow. For independent noise vectors $\epsilon_j \sim \mathcal{N}(0, \mathbf{I}_d)$ and $t \in [0, 1]$, rectified flow (Liu et al., 2023; Lipman et al., 2023) interpolates $I^{(t)} = (1-t)\epsilon + tI$ and learns the target velocity $I - \epsilon$:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{I,c,t,\epsilon}\left[\frac{1}{Nd}\left\|v_{\theta,\phi}(I^{(t)}, t; c) - (I - \epsilon)\right\|_F^2\right]. \tag{1}$$

Sampling integrates the learned field from noise to image latents:

$$\frac{\mathrm{d}\widehat{I}^{(t)}}{\mathrm{d}t} = v_{\theta,\phi}(\widehat{I}^{(t)}, t; c), \qquad \widehat{I}^{(0)} = \epsilon, \qquad I_{\mathrm{gen}} = \widehat{I}^{(1)}. \tag{2}$$

The same construction applies to individual latents or subsets.

3.2 EXISTING UNIFIED MULTIMODAL ARCHITECTURES

The three designs in Figure 1 differ in the image states processed by the backbone and its iteration schedule. We describe them using the rectified-flow formulation above.

Sequence-level diffusion (SLD). SLD jointly denoises the full image, as in Transfusion and Show-o2 (Zhou et al., 2025; Xie et al., 2025). The backbone processes $(c, I^{(t)}, t)$ with bidirectional attention within the noisy image block, and the flow head maps its image states to velocities. Each velocity evaluation refreshes these states, requiring T backbone iterations for T evaluations.

Next-token diffusion (NTD). NTD factorizes $p(I \mid c) = \prod_{j=1}^{N} p(I_j \mid c, I_{<j})$ in a fixed latent order, as in LatentLM and Orthus (Sun et al., 2024; Kou et al., 2025). A causal backbone constructs a condition from each clean prefix; the flow head holds it fixed while generating the next latent. Teacher forcing computes these conditions in parallel, whereas sampling appends each generated latent before predicting the next, requiring N backbone iterations with prefix caching.

Masked autoregressive diffusion (MAD). MAD conditions on visible clean latents, following MAR/MMAR (Li et al., 2024; Yang et al., 2024). An image encoder supplies visible-latent features; the backbone combines them with c and mask embeddings under bidirectional image attention. A per-latent flow head predicts masked positions from these conditions, with loss restricted to masked targets. Sampling holds conditions fixed within each reveal round and recomputes them as generated subsets become visible, requiring K backbone iterations for K rounds.

4 METHOD

CONFLUX uses a content–query (CQ) backbone to construct conditioning representations for text and image prediction (Figure 2).

4.1 SEPARATING SEMANTIC CONDITIONING FROM RENDERING

Given multimodal context c, we place an image query $q_{I,i}$ at each of the N target latent positions. The backbone maps $Q_I = (q_{I,1}, \ldots, q_{I,N})$ to conditions $Z_I = (z_{I,1}, \ldots, z_{I,N})$ for the flow head:

$$Z_I = \left[B_\theta(c; Q_I)\right]_{Q_I}, \qquad v_{\theta,\phi}(I^{(t)}, t; c) = D_\phi(I^{(t)}, t; Z_I). \tag{3}$$

Neither $I^{(t)}$ nor t enters B_θ, so Z_I is computed once per target image and conditioning branch and reused for all evaluations of D_ϕ.

We instantiate D_ϕ as a diffusion transformer (DiT) (Peebles & Xie, 2023; Ma et al., 2024; Esser et al., 2024) with bidirectional self-attention over the full noisy latent grid (Figure 2c). A projection of $z_{I,i}$ is added to a shared time embedding and modulates position i in each layer through adaptive layer normalization (AdaLN). Self-attention couples the latent updates across positions; scaling the head's depth or width leaves the backbone architecture and conditioning interface unchanged.

4.2 Separating Observed Content from Prediction Queries

For text $T = (T_1, \ldots, T_M)$ and image I, the two cross-modal tasks are

$$p(I \mid T), \qquad p(T \mid I) = \prod_j p(T_j \mid I, T_{<j}). \tag{4}$$

Image queries use context $c_I = T$ for T2I; text queries use $c_j = (I, T_{<j})$ for I2T and $c_j = T_{<j}$ for pure-text prediction.

Content and query representations. At position i, content and query states are initialized as $h^0_{c,i} = E(x_i)$ and $h^0_{q,i} = m_{\mathrm{mod}(i)}$. Here E embeds text tokens or projects clean image latents, whereas m is a learned modality embedding independent of the target value. Both streams share attention, normalization, and feed-forward parameters, with one-dimensional RoPE (Su et al., 2021) for text and row/column RoPE for images (Figure 2a).

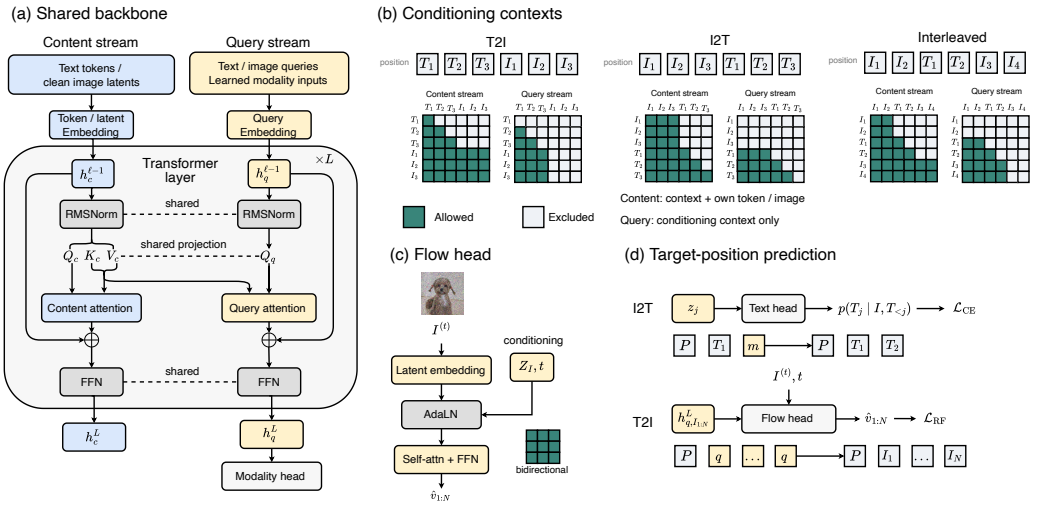


Figure 2: **Content–query factorization and one-pass image conditioning in CONFLUX.** (a) Shared backbone layers update both streams; only content supplies keys and values. (b) Visibility masks for T2I, I2T, and idealized interleaved sequences. Rows index updated states, columns index content states; green marks allowed attention. (c) The flow head processes noisy latents with (Z_I, t) supplied through AdaLN. (d) Image queries condition velocities; text queries parameterize categorical distributions.

Conditioning through content attention. Let y be one text token or one complete image, with preceding context c. Write $H^\ell_c(u)$ and $H^\ell_q(u)$ for the layer-ℓ content and query states at the positions of u. Queries read c, while content additionally reads y:

$$\begin{aligned} \widetilde{H}^\ell_q(y) &= H^\ell_q(y) + \mathrm{Attn}^\ell_\theta\big(H^\ell_q(y), H^\ell_c(c), H^\ell_c(c)\big)\,, \\ \widetilde{H}^\ell_c(y) &= H^\ell_c(y) + \mathrm{Attn}^\ell_\theta\big(H^\ell_c(y), H^\ell_c(c,y), H^\ell_c(c,y)\big)\,. \end{aligned} \tag{5}$$

Here $H^\ell_c(c,y)$ concatenates content states, and Attn includes projections and normalization. Shared residual feed-forward updates then produce $H^{\ell+1}_q$ and $H^{\ell+1}_c$. Only content supplies keys and values

(K/V), excluding query-to-query attention. After L layers and final normalization, query states provide Z_I for the flow head or z_j for the text distribution $p(T_j \mid c_j) = \mathrm{softmax}(W z_j)$ (Figure 2d).

The masks are causal between text tokens and image blocks; within each image, content attention is bidirectional and queries exclude the entire block (Figure 2b). In an idealized teacher-forced pass over $[P, I, S]$, image queries read P to condition $p(I \mid P)$, while image content encodes the complete observed I together with P. Suffix queries read $(P, I, S_{<j})$ to predict S_j, so one backbone pass supplies the conditions for both $p(I \mid P)$ and $p(S_j \mid P, I, S_{<j})$. Since each context's content states read only within that context, queries cannot access their targets indirectly through earlier states (Appendix D.1).

4.3 Joint Training and Generation

Training. We jointly optimize the backbone and heads with text cross-entropy and the rectified-flow loss in Eq. (1):

$$\mathcal{L} = \lambda_{\mathrm{txt}} \mathbb{E}[-\log p(T_j \mid c_j)] + \lambda_{\mathrm{img}} \mathcal{L}_{\mathrm{RF}}. \tag{6}$$

Text-only and I2T examples supervise text queries; T2I examples supervise image velocities, with gradients flowing through the flow head and Z_I to the shared backbone. Each optimizer update accumulates pure-text, T2I, and I2T microbatches, so both objectives train the shared backbone through their respective query states. Appendix A specifies the task mixture and optimization.

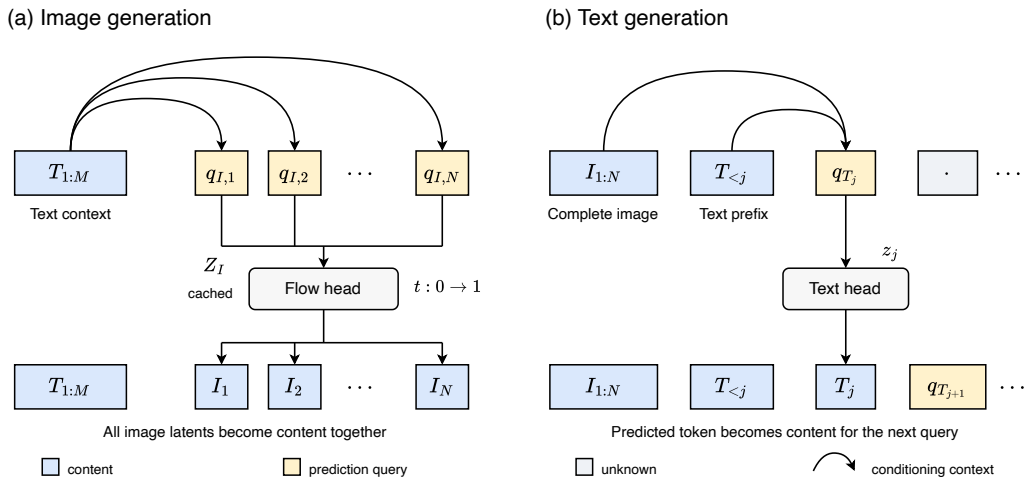


Figure 3: **Conditioning contexts during generation.** Curved arrows connect observed content to prediction queries. (a) Image queries share text context $T_{1:M}$; the flow head generates all latents jointly from fixed Z_I. (b) The text query reads $(I_{1:N}, T_{<j})$ to predict T_j. Lower rows show completed outputs joining the content stream for subsequent predictions.

Generation. Image sampling integrates Eq. (2) with fixed Z_I (Figure 3a). Classifier-free guidance (Ho & Salimans, 2021) uses separate conditions for the conditional and unconditional branches, which can be batched. The completed image enters the content stream; encoding it for later predictions requires a new backbone evaluation. Text generation samples T_j from $p(T_j \mid c_j)$ and adds it to the content stream before predicting T_{j+1} (Figure 3b).

5 Experiments

5.1 Experimental Setup

We compare CONFLUX with SLD, NTD, and MAD (Section 3), initializing all backbones from Qwen3-0.6B-Base (Qwen Team, 2025) and training all their parameters. All configurations are trained with the same random seed. Joint training uses 31,800 updates and 66.7B tokens from ClimbMix text (Diao et al., 2025) and ImageNet-1K image–text pairs (Deng et al., 2009), with shared

data order and optimization. All models use rectified flow with a frozen KL-16 autoencoder (Li et al., 2024) for 256×256 images.

All models receive identical complete, clean I2T inputs under their native attention rules. Text-only controls use the same pure-text stream and corresponding readout. Final evaluations use EMA weights; language and understanding summaries average eight and five benchmarks, respectively. Generation quality uses FID-selected settings, and throughput uses common sampling settings. Appendices A–C provide training, evaluation, and architecture details and benchmark breakdowns.

5.2 Language and Visual Understanding

CONFLUX improves both language and visual understanding (Table 1). It exceeds the strongest respective baselines by 7.56 points on MMLU and 2.38 points on language macro average, and leads on four of five understanding benchmarks, with gains of 3.53 points on ImageNet classification and 1.96 points on SugarCrepe.

Table 1: **Language and visual understanding.** NTP and TP denote next-token and target-position text readouts. Bold marks the best multimodal score in each column.

	Language		Understanding				
Model	Macro ↑	MMLU ↑	IN-1K ↑	SugarCr. ↑	ARO-R ↑	MMB ↑	SEED ↑
Text-only, NTP	57.03	44.94	—	—	—	—	—
Text-only, TP	57.21	45.38	—	—	—	—	—
Iterative-backbone baselines							
SLD	48.29	27.54	43.06	77.39	70.65	43.30	39.51
NTD	49.30	28.93	36.63	75.06	70.39	38.49	36.97
MAD	52.29	35.01	41.57	76.46	**72.05**	37.97	34.93
CONFLUX	**54.67**	**42.57**	**46.59**	**79.35**	71.75	**44.67**	**40.64**

The language-retention advantage persists throughout joint training (Figure 4). At approximately 10B pure-text tokens, CONFLUX trails its text-only control by 1.0 MMLU point, versus 6.1–11.7 points for the baselines. At 33.3B tokens, the gaps are 3.3 and 9.9–17.0 points, respectively.

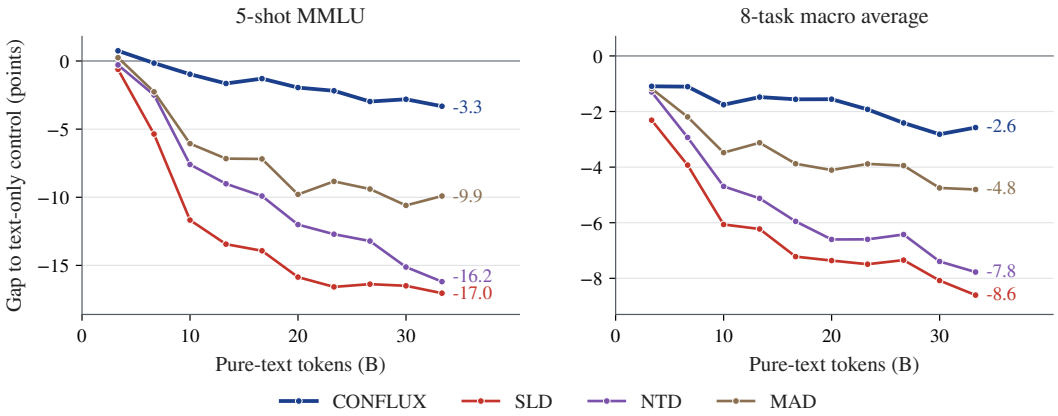


Figure 4: **Language retention throughout joint training.** Scores are relative to the corresponding text-only controls; zero denotes no gap. Curves use a common training-time evaluation harness; Table 1 reports final EMA scores.

5.3 Image Generation and Flow-Head Scaling

CONFLUX achieves $4.0\times$ the throughput of SLD, the fastest baseline, with FID 3.67 versus 3.09 (Table 2).

Table 2: **Image generation and flow-head scaling.** Parentheses give the flow-head size. Throughput uses common sampling settings; FID and IS use FID-selected settings.

		Average score (%)		Generation		
Model	Total (B)	Language ↑	Understanding ↑	FID ↓	IS ↑	Images/s ↑
SLD	0.76	48.29	54.78	**3.09**	**220.2**	71.1
NTD	0.76	49.30	51.51	4.98	183.2	24.0
MAD	0.97	52.29	52.60	5.03	200.6	33.7
CONFLUX (0.16B)	0.76	54.67	56.60	3.67	169.8	**287.0**
CONFLUX (0.33B)	0.93	**54.95**	**57.39**	3.25	188.9	177.4
CONFLUX (0.67B)	1.27	54.20	57.21	3.19	202.3	115.6

Scaling only the flow head further reduces FID to 3.19 and raises IS from 169.8 to 202.3. Across head sizes, average language and understanding scores remain within 0.5 and 0.8 points of the default, respectively, while the largest head retains a $1.6\times$ throughput advantage over SLD. The backbone architecture and conditioning schedule are unchanged across these runs, placing the additional capacity in the iterative renderer.

5.4 Architectural Ablations

CONFLUX achieves higher language and understanding scores and lower FID than its AR and masked AR variants, although these variants attain higher IS (Table 3). Relative to AR, strict AR reduces understanding by 2.32 points with little change in FID. Raster order is more detrimental: FID rises from 4.28 to 17.42, and both language and understanding decline.

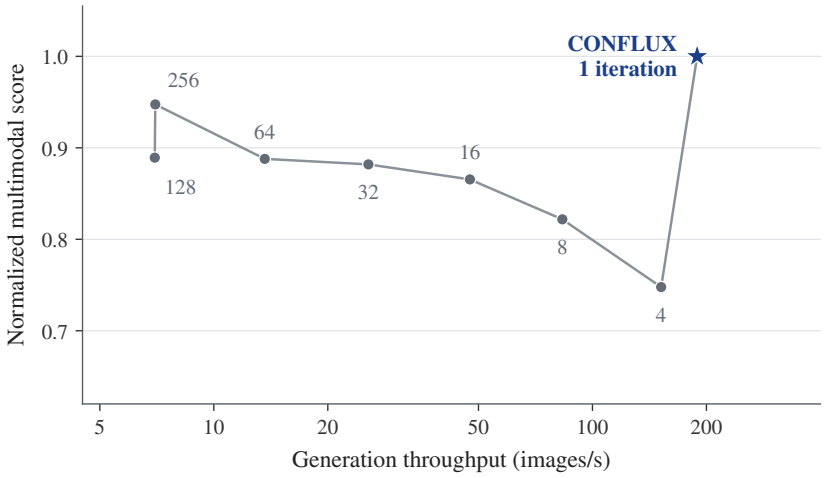


Figure 5: **Quality–speed trade-off across CONFLUX variants.** Labels give backbone iterations per image and conditioning branch: 1 for the default (star), 4–128 for masked AR, and 256 for AR. The score averages normalized language, understanding, and inverse FID (Appendix C.4).

Reducing backbone iterations accelerates masked AR at the expense of FID (Figure 5). The default one-pass variant achieves both the highest normalized multimodal score and throughput.

Text CQ raises CONFLUX's language and understanding averages by 0.86 and 0.68 points, respectively, with little change in FID (Appendix D.2).

Table 3: **Architectural variants.** AR predicts individual latents; masked AR predicts subsets. Strict AR excludes the content-attention diagonal; raster AR uses raster order.

	Average score (%)		Generation	
Variant	Language ↑	Understanding ↑	FID ↓	IS ↑
CONFLUX	54.67	56.60	3.67	169.8
CONFLUX, AR	52.30	54.09	4.28	201.8
CONFLUX, masked AR	52.20	50.56	5.20	186.3
CONFLUX, strict AR	52.03	51.77	4.20	224.0
CONFLUX, raster AR	48.65	50.00	17.42	79.8

5.5 Cross-Modal Representational Alignment

We independently encode 1,000 ImageNet object–name pairs and mean-pool content states within each modality. The mNN@10 score measures agreement between their local neighborhoods, testing whether the two modalities organize paired examples similarly.

CONFLUX achieves the highest best-pair and final-layer alignment among the four models (Figure 6). Relative to SLD, the strongest baseline on both measures, best-pair mNN@10 rises from 0.0566 to 0.0821 (+45.1%), and final-layer mNN@10 from 0.0259 to 0.0636 (2.46×). Even CONFLUX's final-layer score exceeds every baseline's best-pair score.

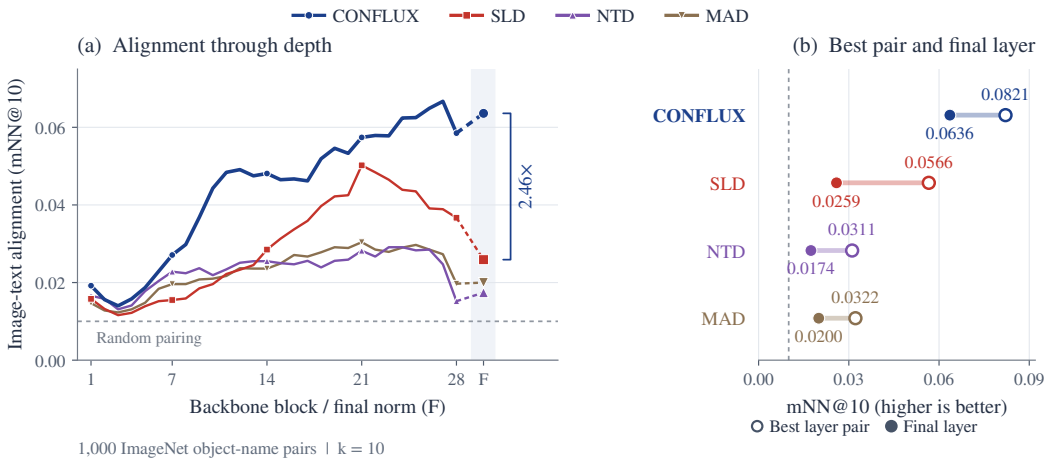


Figure 6: **Cross-modal alignment on ImageNet object–name pairs.** mNN@10 measures overlap between image and text 10-neighbor sets (Huh et al., 2024). **(a)** Corresponding-layer scores; F denotes final RMSNorm. **(b)** Best-pair (open) and final-layer (filled) scores. Dashed lines mark chance-level overlap (Appendix D.3).

6 Conclusion

CONFLUX separates semantic conditioning from iterative rendering and observed content from prediction queries. Its shared content–query interface supports text and image prediction while preserving observed context, and its flow head reuses image conditions computed in a single backbone pass. Under matched data and update budgets, CONFLUX retains more language capability, improves most visual understanding metrics, and increases image-generation throughput over iterative-backbone baselines. The learned representations also show stronger cross-modal neighborhood alignment. Scaling the flow head further improves generation quality while largely preserving language and visual understanding performance, without changing backbone capacity or the one-pass conditioning schedule.

AI USE STATEMENT

AI assistance was used in restructuring and drafting this manuscript, clarifying its conceptual framing and attention-mask distinctions, proposing evaluation controls, and preparing LaTeX and figure code. The primary experimental results originate from the authors' runs. AI assistance also supported execution of the fixed-subset evaluations and matched timing diagnostic in Appendix C.4, using the authors' checkpoints and evaluation code; the measured outputs and protocols were retained. Authors remain responsible for the manuscript, verification of claims and references, and disclosure of any additional AI use.

REPRODUCIBILITY STATEMENT

Section 4 specifies the content–query masks and generation schedule. Appendix A records training and evaluation settings. Appendices B and C provide benchmark breakdowns, sampling settings, and flow-head configurations. Appendix C.4 specifies the quality–speed diagnostic, and Appendix D.1 explains the no-leakage property. Exact implementation revisions and selected generation settings should accompany the experimental release.

REFERENCES

Victor Besnier, Mickael Chen, David Hurych, Eduardo Valle, and Matthieu Cord. Halton scheduler for masked generative image transformer. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.17076.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2020. arXiv:1911.11641.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. URL https://aclanthology.org/N19-1300/.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan Lin, Jan Kautz, and Pavlo Molchanov. Nemotron-CLIMB: CLustering-based iterative data mixture bootstrapping for language model pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.13161.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.11499.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.03206.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2023. URL https://zenodo.org/records/10256836.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2009.03300.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.08500.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. arXiv:2207.12598.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14610.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2412.00127.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11838.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/lin24c.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.03003.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2307.06281.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2401.08740.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. URL https://aclanthology.org/D18-1260/.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2212.09748.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence*, 2020. arXiv:1907.10641.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1606.03498.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024.

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 arXiv:2601.16208*, 2026. URL https://arxiv.org/abs/2601.16208.

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

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.15564.

Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. MMAR: Towards lossless multi-modal auto-regressive probabilistic modeling. *arXiv preprint arXiv:2410.10798*, 2024.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1906.08237.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2001.06782.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.01936.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. URL https://aclanthology.org/P19-1472/.

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

A Training and Evaluation Details

A.1 Training Configuration

Table 4: **Training configuration.** Learning rates are peak values. Flow time is shared across the latent positions of each image.

Setting	Value
Backbone	Qwen3-0.6B-Base, fully trainable
Image autoencoder	Frozen KL-16; 256×256 images
Image representation	256 latents, 16 channels each
Devices / optimizer updates	64 Ascend 910B / 31,800
Backbone and text-embedding learning rate	3×10^{-4}
Image-projection and flow-head learning rate	5×10^{-4}
Text / image loss weight	0.05 / 1.0
Flow-time distribution	90% logit-normal $(0, 1)$ + 10% uniform
Image-content augmentation (CONFLUX)	Gaussian noise, standard deviation 0.01
Image RoPE base	10^4
Final evaluation weights	EMA
Training seeds per configuration	1

Table 5: **Composition of the training budget.** Each update accumulates two pure-text microbatches, one T2I microbatch, and one I2T microbatch. Counts include padding. Image latents are prediction targets in T2I and observed inputs in I2T.

Stream	Sequences per device/microbatch	Length	Total positions	Image latents
Pure text	4	2048	33.3B	–
T2I	16	512	16.7B	8.3B
I2T	16	512	16.7B	8.3B
Total			66.7B	16.7B

Comparisons match data exposure and optimizer updates, rather than training FLOPs or total parameter count. Text-only controls use the same pure-text stream and order. Image-content augmentation in CONFLUX is independent of flow time, leaves rectified-flow targets unchanged, and is disabled at evaluation. Image losses average squared velocity errors over latent positions and channels.

Gradient-projection methods such as PCGrad (Yu et al., 2020) address task interference at the optimizer level; the primary comparison varies architecture under a common optimizer.

A.2 Evaluation Protocol

Language evaluation follows the task prompts, splits, and metrics of the Language Model Evaluation Harness (Gao et al., 2023), with seven zero-shot tasks and 5-shot MMLU (Hendrycks et al., 2021). ARC-easy/challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), and OpenBookQA (Mihaylov et al., 2018) normalize likelihood by the original answer choice's character count; WinoGrande (Sakaguchi et al., 2020) and BoolQ (Clark et al., 2019) use accuracy. MMLU averages accuracy over 57 subjects (14,042 questions). The language macro score averages the eight task scores.

Figure 4 uses a common training-time evaluation harness for all checkpoints and text-only controls. Its endpoint gaps are computed within that evaluation series and differ from the final EMA evaluations in Table 1.

Visual understanding ranks candidate texts by mean token log-likelihood $s(I, y)$, with language-prior correction motivated by VisualGPTScore (Lin et al., 2024). ImageNet classification, SugarCrepe (Hsieh et al., 2023), MMBench (Liu et al., 2024), and SEED-Bench (Li et al., 2023) use the debiased score $s(I, y) - \log[|\mathcal{R}|^{-1} \sum_{J \in \mathcal{R}} \exp s(J, y)]$, with prior weight one. The reference set

$\mathcal{R}$ contains all 50K evaluation images for ImageNet and three fixed Gaussian images for the other tasks. ARO-relation (Yuksekgonul et al., 2023) uses $s(I, y)$ directly. MMBench requires correctness across all option rotations. FID (Heusel et al., 2017) and IS (Salimans et al., 2016) use 50K generated images; FID uses ImageNet validation statistics. Generation settings are selected from FID sweeps on the validation distribution used for reporting; MAD uses 32 reveal rounds (FID 5.034), with near-identical FID at 256 rounds (5.032).

A.3 Generation Throughput

Table 6: **Throughput protocol for Table 2.** Conditional and unconditional branches are batched. Measurement covers latent generation and excludes VAE decoding and metric computation. FID-selected settings are listed separately in Table 10.

Setting	Value
Devices	16 Ascend 910B
Images per device before guidance doubling	64
Classifier-free guidance	3.5
Solver	Heun, 10 steps (20 velocity evaluations)
Model / integration precision	BF16 / FP32
Warm-up / measured batches	2 / 5
Measured images per configuration	5,120
Prompts and initial noise	Identical across models
MAD reveal rounds / order	32 / spatial Halton

Spatial Halton ordering follows Besnier et al. (2025). Backbone iterations count the construction of image conditions per target image and conditioning branch. Prompt prefill and query computation may be fused or use an equivalent content cache. Flow-head iterations, text continuation, and subsequent encoding of generated images as context are excluded from this count. The primary throughput comparison uses common sampling settings; generation quality uses separately selected settings, so the reported speed ratios are not iso-quality measurements.

B Language and Visual Understanding Details

B.1 Language Benchmark Breakdown

Table 7: **Per-benchmark language scores (%).** Full evaluations use final EMA checkpoints; Macro averages the eight tasks. ARC-e/c denote ARC-easy/challenge, HSw denotes HellaSwag, WG denotes WinoGrande, and OBQA denotes OpenBookQA.

Model	ARC-e ↑	ARC-c ↑	HSw ↑	PIQA ↑	WG ↑	BoolQ ↑	OBQA ↑	MMLU ↑	Macro ↑
SLD	61.03	31.74	48.25	70.51	55.09	57.80	34.40	27.54	48.29
NTD	60.94	32.34	48.03	70.73	55.64	61.62	36.20	28.93	49.30
MAD	65.11	35.58	53.45	72.25	57.70	63.21	36.00	35.01	52.29
CONFLUX (0.16B)	67.38	38.57	57.71	73.45	57.54	61.74	38.40	42.57	54.67
CONFLUX (0.33B)	67.09	37.20	57.90	73.23	59.59	64.50	37.80	42.33	54.95
CONFLUX (0.67B)	66.50	39.16	57.69	73.50	58.64	58.26	36.60	43.25	54.20
CONFLUX, w/o Text CQ	66.67	36.60	56.19	72.85	58.33	62.08	36.40	41.34	53.81
SLD, Text CQ	61.07	35.15	50.23	71.76	55.33	60.43	34.20	29.72	49.74
NTD, Text CQ	60.90	34.04	49.17	71.44	56.99	59.60	33.40	30.96	49.56
MAD, Text CQ	63.97	35.92	54.56	71.60	57.14	58.87	36.60	37.28	51.99

B.2 VISUAL UNDERSTANDING BREAKDOWN

Table 8: **Category-level visual understanding (%).** All categories are shown for SugarCrepe, MMBench, and SEED-Bench; ARO-relation shows three selected spatial relations. N counts evaluated examples, or original questions for MMBench before option rotation. Scores use each benchmark's primary metric.

Category	N	SLD ↑	NTD ↑	MAD ↑	CONFLUX ↑
SugarCrepe					
Add attribute	692	76.45	72.69	77.02	**78.32**
Add object	2,062	82.06	78.37	79.83	**83.41**
Replace attribute	788	80.33	80.71	79.95	**82.36**
Replace object	1,652	79.36	76.21	78.81	**83.11**
Replace relation	1,406	67.21	66.93	64.58	**69.27**
Swap attribute	666	78.83	77.33	**81.08**	79.13
Swap object	245	72.65	68.57	**75.10**	71.43
ARO-relation: selected spatial relations					
To the left of	7,741	65.30	65.46	64.80	**68.29**
To the right of	7,741	66.62	66.97	70.35	**71.15**
In front of	588	74.49	**75.51**	68.54	51.53
MMBench					
Attribute reasoning	199	40.70	36.68	33.17	**42.21**
Coarse perception	296	**57.43**	55.07	48.99	**57.43**
Fine-grained perception: cross-instance	143	49.65	39.86	43.36	**55.94**
Fine-grained perception: instance-level	293	37.88	32.76	32.08	**39.25**
Logic reasoning	118	25.42	22.88	25.42	**29.66**
Relation reasoning	115	35.65	27.83	**39.13**	31.30
SEED-Bench					
Instance Attributes	4,649	**40.98**	38.14	38.46	38.65
Instance Identity	1,831	29.49	30.31	29.77	**33.81**
Instance Interaction	97	**47.42**	39.18	35.05	45.36
Instance Location	978	37.53	36.50	33.84	**39.78**
Instances Counting	2,447	36.29	32.12	27.58	**37.02**
Scene Understanding	3,158	47.21	43.76	39.61	**50.76**
Spatial Relation	657	33.33	29.98	29.83	**37.14**
Text Understanding	85	43.53	**55.29**	50.59	34.12
Visual Reasoning	331	39.58	38.37	32.63	**46.22**

Selected correct predictions appear in Appendix E.1.

C Separating Semantic Conditioning from Rendering

C.1 Generation Architectures and Sampling Settings

Table 9: **Generation architectures in the primary comparison.** All use continuous image latents and rectified flow. MAD additionally uses a 16-layer visible-latent encoder.

Model	Backbone conditioning	Flow head
SLD	Noisy image states, updated at each velocity evaluation.	8-layer, width-1280 transformer; projected backbone states and flow time.
NTD	Causal image content; one condition per latent, fixed within its flow solve.	8-layer, width-1280 transformer; noisy latent queries attend to committed latents.
MAD	Visible content and placeholders, updated after each reveal round.	8-block, width-1936 per-position MLP; noisy latents conditioned on placeholder states.
CONFLUX	Learned image queries read text content once.	12-layer, width-864 DiT; noisy latents with fixed backbone conditions and flow time via AdaLN.

Table 10: **Sampling settings for the reported generation scores.** All rows use Heun10 and 50K samples from final EMA checkpoints. Backbone iterations are per image and conditioning branch. Joint generation completes all latents together; raster and Halton specify latent completion order. FID and IS use the same operating point.

Model	CFG	Backbone iterations	Order	FID ↓	IS ↑
SLD	1.5	20	Joint	3.09	220.2
NTD	3.0	256	Raster	4.98	183.2
MAD	4.5	32	Halton	5.03	200.6
CONFLUX (0.16B)	1.5	1	Joint	3.67	169.8
CONFLUX (0.33B)	1.5	1	Joint	3.25	188.9
CONFLUX (0.67B)	1.5	1	Joint	3.19	202.3
CONFLUX, w/o Text CQ	1.5	1	Joint	3.71	167.9
SLD, Text CQ	1.5	20	Joint	3.03	242.1
NTD, Text CQ	3.0	256	Raster	5.07	181.1
MAD, Text CQ	3.0	32	Halton	4.54	190.3

C.2 Flow-Head Scaling

Table 11: **Flow-head architectures.** All three configurations use the same 0.6B backbone architecture and an attention-head dimension of 72. Flow-head parameter counts use the rounded sizes in Table 2.

Flow head	Layers	Width	Attention heads	FFN width
0.16B	12	864	12	3456
0.33B	18	1008	14	4032
0.67B	28	1152	16	4608

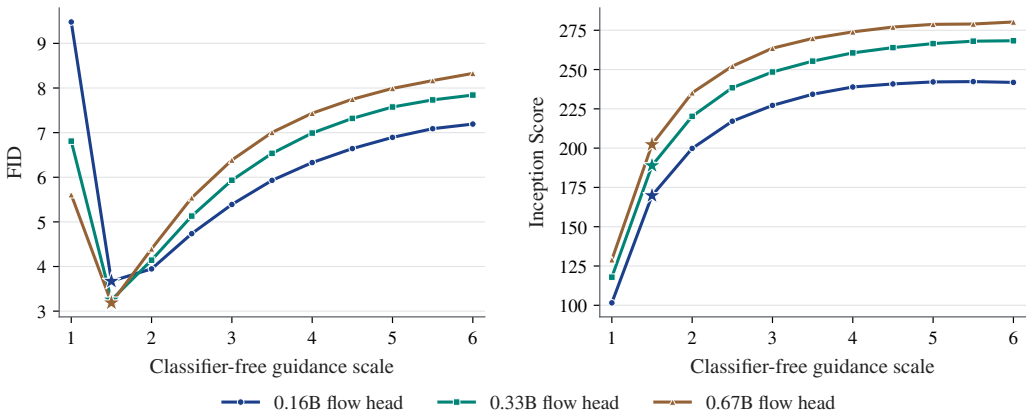


Figure 7: **Guidance sweeps for flow-head scaling.** Each point uses 50K samples from the corresponding final EMA checkpoint, with Heun10 and seed 42. Stars mark each checkpoint's FID-selected setting on both panels.

C.3 Architectural Variants

Table 3 retains the content–query text interface across all variants. AR uses random-order latent autoregression, causal image content, and an 8-layer, width-1280 contextual flow head. Masked AR predicts unknown latents from a visible subset during T2I, retains complete bidirectional image content for I2T, and uses an 8-block, width-1936 per-position flow head. Strict AR excludes the content-attention diagonal in both backbone and flow head; query visibility is unchanged. Raster AR replaces random-order training and Halton sampling with raster order. Comparisons across generation factorizations include their associated flow heads.

Generation uses 50K images, Heun10, and seed 42. CFG is 1.5 for CONFLUX, 4.5 for masked AR, and 2.0 for AR, strict AR, and raster AR. Masked AR uses 256 reveal rounds after a CFG sweep at 64 rounds; AR and strict AR use Halton order, and raster AR uses raster order. Language averages eight full benchmarks; understanding averages the same five benchmarks as Table 2, using native likelihood estimators: 64 Monte Carlo samples for random-order AR variants and one for the joint, masked, and raster variants on the compositional and multiple-choice tasks.

C.4 Quality–Speed Diagnostic

For Figure 5, let T_m, U_m, and F_m denote language score, visual understanding score, and FID. With the default one-pass variant as reference r, the equally weighted summary is

$$Q_m = \frac{1}{3}\left(\frac{T_m}{T_r} + \frac{U_m}{U_r} + \frac{F_r}{F_m}\right). \tag{7}$$

Language averages the eight tasks on fixed subsets of 128 examples each; MMLU covers all 57 subjects with 5-shot prompts. Understanding averages SugarCrepe, ARO-relation, and SEED-Bench on 256 examples each, using one image-likelihood sample per example. Subsets use seed 20260926. FID uses the full 50K-sample evaluations.

Table 12: **Sampling settings for Figure 5.** All settings use Heun10. The reveal-count sweep shares one checkpoint and therefore shares its language and understanding scores. CFG is selected at 64 iterations and held fixed across that sweep.

Backbone iterations	CFG	Generation schedule
1	1.5	Joint, fixed image conditions
4, 8, 16, 32, 64, 128	4.5	Random reveal order
256	2.0	Spatial Halton order, backbone caching

Throughput uses 16 Ascend 910B devices, eight images per device, identical prompts and initial noise, BF16 models, and FP32 integration. After one warm-up batch, throughput divides the global batch size by the median of two synchronized batch times, each determined by the slowest device. Measurement excludes input transfer, VAE decoding, and metric computation; its batch size differs from Table 2.

The variants differ in flow-head architecture and conditioning schedule as well as backbone iteration count. The reveal-count checkpoint also uses a 5×10^{-5} flow-head and image-projector learning rate, versus 5×10^{-4} for the one- and 256-iteration variants. Guidance is not optimized separately at each reveal count. This diagnostic therefore compares the evaluated configurations rather than isolating iteration count alone.

Selected generation examples and matched-noise flow-head comparisons appear in Appendices E.2 and E.3, respectively.

D Separating Observed Content from Prediction Queries

D.1 Content Visibility

No target leakage through content. Assume that Eq. (5) defines the only cross-position operations in each layer, with position-wise normalization and feed-forward updates. Fix the model parameters, positions, and any independent input-augmentation noise. Consider a target y, consisting of one text token or one complete image, and its preceding context c. At initialization, $H_c^0(c)$ depends only on the observed values in c, while $H_q^0(y)$ depends only on modality embeddings and target positions. For every token or image contained in c, both that token or image and its own preceding context are contained in c. Its content attention therefore reads only content states within c. By induction over layers, $H_c^\ell(c)$ cannot depend on y or later values. The query update reads only these states and the query's previous state, so $H_q^L(y)$ also excludes the target values. In contrast, $H_c^L(y)$ may encode the complete observed y for subsequent predictions.

Two roles in an idealized mixed sequence. For an idealized sequence $[P, I, S]$, teacher forcing can compute the conditioning representations for $p(I \mid P)$ and all text factors $p(S_j \mid P, I, S_{<j})$ in one backbone pass. At layer ℓ, image queries read $H_c^\ell(P)$, whereas image content reads $H_c^\ell(P, I)$. Response queries read $H_c^\ell(P, I, S_{<j})$. Thus separate states represent the image as a prediction target and as complete observed context. The evaluated task mixture uses separate T2I and I2T microbatches; no mixed-sequence benchmark is reported here.

D.2 Text-Interface Comparisons

We compare target-position (TP) text queries with next-token prediction (NTP) for each image-generation architecture. With Text CQ, CONFLUX gains 0.86 and 0.68 points in average language and understanding, with similar FID (Table 13). SLD and NTD also score higher on both averages, while MAD scores lower.

Table 13: **Text CQ across image-generation architectures.** TP uses target-position queries (Text CQ); NTP uses single-stream next-token prediction for text and I2T. Language and understanding average eight and five benchmarks, respectively. FID and IS use FID-selected settings.

		Average score (%)		Generation	
Image design	Text	Language ↑	Understanding ↑	FID ↓	IS ↑
SLD	NTP	48.29	54.78	3.09	220.2
	TP	49.74	55.66	3.03	242.1
NTD	NTP	49.30	51.51	4.98	183.2
	TP	49.56	51.79	5.07	181.1
MAD	NTP	52.29	52.60	5.03	200.6
	TP	51.99	51.38	4.54	190.3
CONFLUX	NTP	53.81	55.92	3.71	167.9
	TP	54.67	56.60	3.67	169.8

D.3 Cross-Modal Representational Alignment

Samples and independent inputs. We adapt the mutual nearest-neighbor diagnostic of Huh et al. (2024) to 1,000 ImageNet-1K validation object–name pairs, selecting one image per class with seed 42. For each image, we crop the largest official localization box matching its ground-truth class and pair it with the canonical English class name. Crops preserve aspect ratio, fit within 256×256 pixels, and receive centered gray padding. The frozen MAR KL-16 autoencoder supplies one fixed posterior draw per image, shared across the four models. Each backbone encodes the image and class name in separate, unprompted forward passes, with all clean image latents supplied and native understanding attention retained. Thus the diagnostic measures the geometry of observed content, not image-generation queries. All four checkpoints use final EMA weights at 31,800 updates.

Features and metric. We average content-token states within each modality, excluding boundary and padding tokens. Features are collected at the embedding, each of the 28 Transformer blocks, and the final RMSNorm output, before any next-token readout shift. Per modality, we clip pooled feature coordinates at the global 95th percentile of their absolute values across samples, layers, and channels, then apply per-sample ℓ_2 normalization. For layer choices a and b, let $\mathcal{N}^k_{I,a}(i)$ and $\mathcal{N}^k_{T,b}(i)$ denote the cosine-similarity neighbor indices of paired sample i in the image and text spaces, respectively. Excluding the sample itself, we compute

$$A_{a,b} = \frac{1}{Nk} \sum_{i=1}^{N} \left| \mathcal{N}^k_{I,a}(i) \cap \mathcal{N}^k_{T,b}(i) \right|, \qquad N = 1{,}000, \quad k = 10. \tag{8}$$

Neighbors are selected from the complete sample set, independently of extraction batch size. Under random pairing, expected overlap is $k/(N-1) = 10/999$.

Layer selection and interpretation. The best-pair score maximizes $A_{a,b}$ over blocks 1–28, final RMSNorm, and the concatenation of all block features, excluding embeddings. We also report corresponding-layer scores and the fixed final-RMSNorm pair to expose differences between intermediate and output geometry. The same sample set is used for layer selection and reporting, so the maxima are descriptive. The original PRH experiment uses 1,024 WIT image–caption pairs and image CLS features; our object crops, class names, and image-token pooling define a different diagnostic. Absolute scores are therefore specific to this protocol. The configurations differ in their prediction interfaces and retain their native attention patterns; one checkpoint per configuration cannot establish training-seed uncertainty or attribute alignment gains to an individual mechanism.

E Qualitative Examples and Failure Cases

E.1 Understanding Examples

(a) Occupation recognition | MMBench #1398



What's the profession of the people in this picture?

A. driver
B. teacher
C. janitor
D. tailor

SLD	NTD	MAD	CONFLUX
B	A	A	C

(b) Size and relation binding | SugarCrepe swap_att/628



A. a small elephant walks underneath a big elephant
B. a big elephant walks underneath a small elephant.

SLD	NTD	MAD	CONFLUX
B	B	A	A

(c) Counting visible objects | SEED-Bench #38507



How many chairs are present in the image?

A. Two
B. One
C. Three
D. Four

SLD	NTD	MAD	CONFLUX
C	A	C	A

Figure 8: **Selected correct predictions by CONFLUX.** Images are the evaluated 256×256 crops. Each model's highest-scoring candidate is marked green if correct and red otherwise. MMBench shows the original option order; its aggregate score requires correctness under all rotations.

E.2 Selected Generation Examples

All examples use final EMA weights and 256×256 resolution. Each row shares a caption across models; sampling settings and seeds may differ. Table 2 reports aggregate generation performance.



Figure 9: **Selected generation examples (1/3).** Each row shares a caption across models.

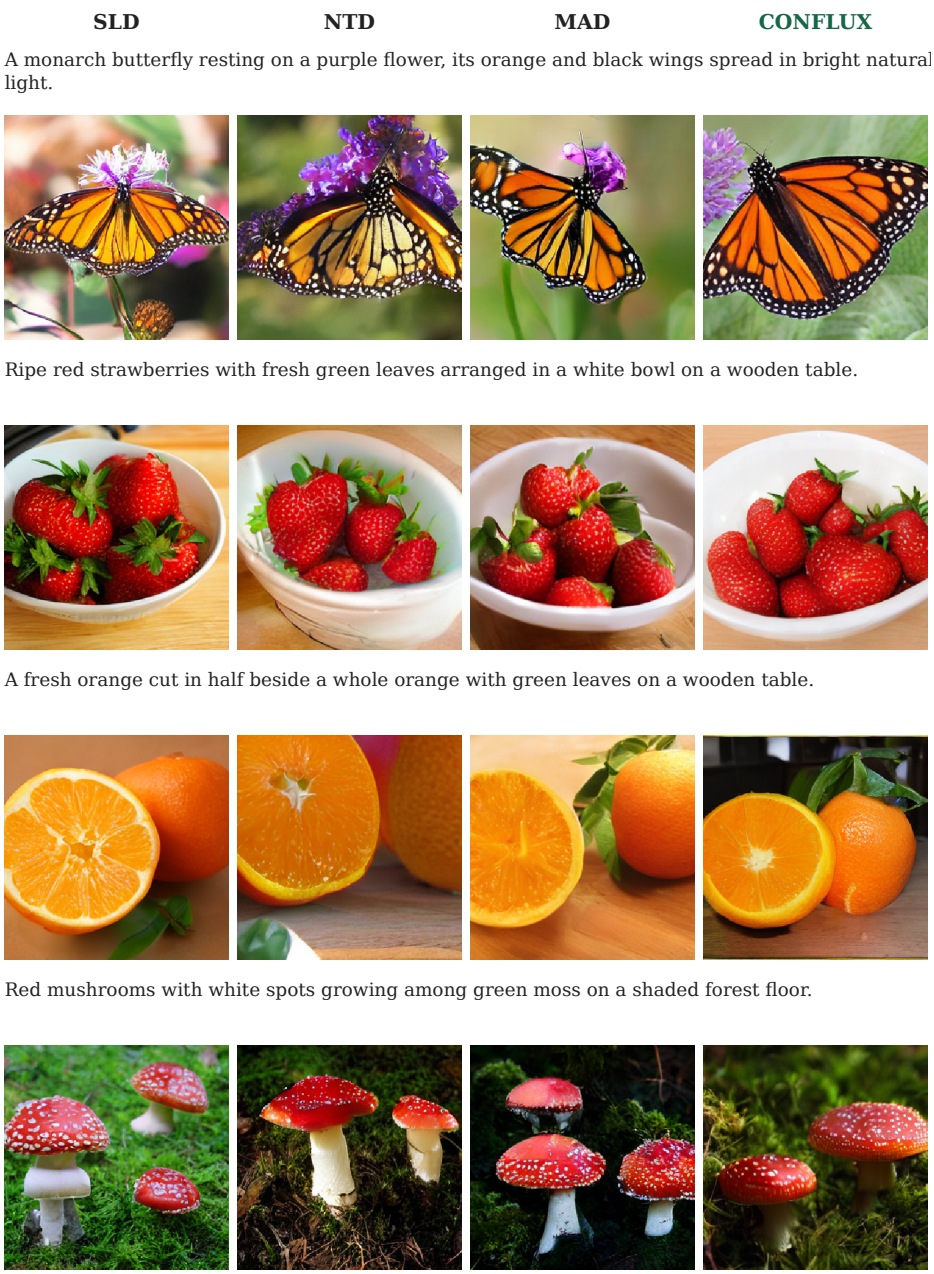


Figure 10: **Selected generation examples (2/3).** Each row shares a caption across models.



Figure 11: **Selected generation examples (3/3).** Each row shares a caption across models.

E.3 Flow-Head Scaling Examples

0.16B flow head **0.33B flow head** **0.67B flow head**

A red dahlia flower with many layers of neatly arranged petals, photographed close up in a green garden.



A small timber-framed house with a tiled roof and flower boxes beneath its windows, in a green garden.



Figure 12: **Selected flow-head scaling examples (1/2).** Each row uses the same caption and initial noise across the 0.16B, 0.33B, and 0.67B flow heads, with CFG 1.5, Heun10, and 256×256 outputs. Table 2 reports aggregate performance.

0.16B flow head **0.33B flow head** **0.67B flow head**

A Gothic stone church with tall spires, pointed arches and a large rose window in clear daylight.



A red vintage tram on a cobblestone street, with tall windows and overhead wires in warm afternoon light.



Figure 13: **Selected flow-head scaling examples (2/2).** Each row uses the same caption and initial noise across the 0.16B, 0.33B, and 0.67B flow heads, with CFG 1.5, Heun10, and 256×256 outputs. Table 2 reports aggregate performance.

E.4 Representative Failure Cases

(a) Image generation: extraneous detail



Prompt: A green tree frog sitting sideways on a leaf, with its front legs, round toe pads and folded hind legs clearly visible.

CFG 1.75; Heun20; temperature 1.0; seed 1022.

Analysis: The frog and leaf are recognizable, but unrequested text-like marks appear near the lower edge. This example preserves the requested subject while introducing extraneous visual detail.

(b) Captioning: invented object and action



Prompt: Describe this image in one detailed caption:

CONFLUX excerpt: "stands at bat, gripping a wooden bat mid-swing."

Analysis: The image shows a player holding a ball and glove, preparing to throw. The caption recognizes the baseball scene but substitutes batting for throwing and introduces an absent bat.

(c) Multiple-choice understanding: counting



Question: How many people are playing music in the image?

A. One | B. Two | C. Three | D. Four

CONFLUX: B. Two

Reference: C. Three

Analysis: Three musicians are visible: two front performers and a drummer. CONFLUX selects two, undercounting the group by one.

Figure 14: **One failure case per task.** All outputs use the final EMA CONFLUX checkpoint. **(a)** Original generated image. **(b)** VAE reconstruction of the encoded captioning input, with a verbatim excerpt from greedy decoding. **(c)** Evaluated SEED-Bench crop, with the recorded prediction and reference answer; the decision uses language-prior-debiased likelihood.

F Limitations

Evaluation is limited to one backbone family and ImageNet image–text training at 256×256 resolution. Generalization to broader visual data, larger backbones, and higher resolutions remains untested. Language performance remains below the matched text-only control, and the mixed-sequence visibility advantage has not been evaluated on an interleaved benchmark.