EXPLORING UNIFIED LATENT DIFFUSION FOR MULTIMODAL MODELING

Anonymous authors
Paper under double-blind review

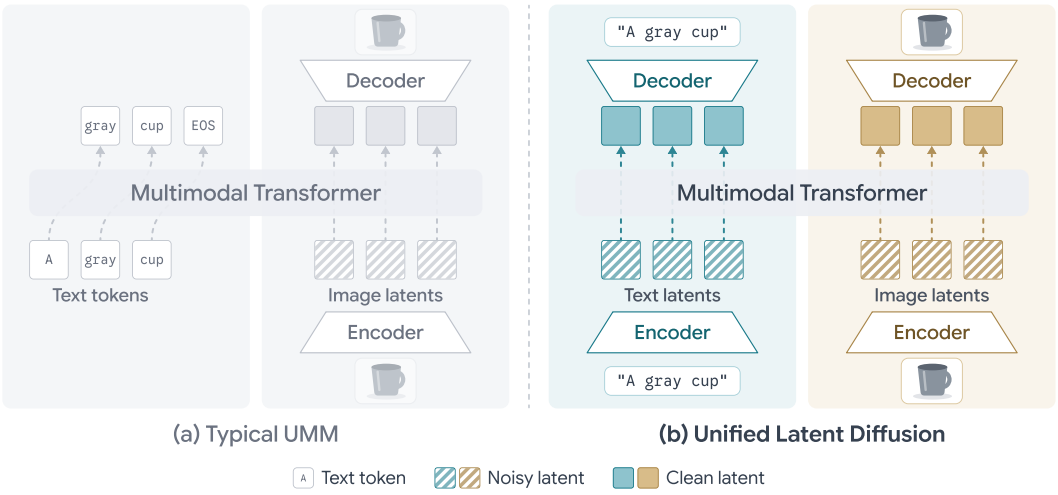

Figure 1: **Unified Latent Diffusion** generates both modalities through continuous latent denoising, unifying multimodal generative modeling. Typical UMMs use different generative modeling for language (autoregressive next-token prediction) and vision (continuous latent denoising).

ABSTRACT

Unified multimodal models commonly combine autoregressive language generation with continuous image diffusion, requiring a shared model to learn two different generative processes. We study **Unified Latent Diffusion**, which generates both language and images through continuous latent denoising, and ask how this choice affects joint multimodal learning. Through controlled pretraining from scratch, we compare autoregressive, discrete diffusion, and latent diffusion language modeling with matched backbone architectures, training data, and token budgets, while keeping visual modeling fixed. We identify three findings: (i) Unifying generative modeling through latent diffusion eases joint training, surpassing the autoregressive baseline's visual generation (DPG-Bench) with one-third fewer training tokens. (ii) Iterative denoising refines answers and strengthens visual understanding (Vision-Centric: 33.76 vs. 20.84 for AR at 30B training tokens). (iii) Effective language latents for diffusion modeling require different properties from vision. These comparisons examine the capabilities and limitations of unified latent diffusion as a generative formulation for general-purpose UMMs.

1 INTRODUCTION

Generative modeling has become a foundation of modern AI, but the dominant methods differ by modality. Language generation primarily uses autoregressive modeling (Brown et al., 2020; Llama, 2023; Yang et al., 2025), while visual generation commonly relies on continuous latent diffusion (Rombach et al., 2022; Esser et al., 2024; Labs et al., 2025). Research on general-purpose generative models is now directed toward unified multimodal models (UMMs), which generate both vision and language within a single model (Tong et al., 2026a). The typical strategy for UMMs is

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.

to combine autoregressive language models with visual latent diffusion models (Zhou et al., 2025; Deng et al., 2025). This approach combines two distinct prediction objectives: discrete next-token prediction for language and continuous latent denoising for vision. However, these modality-specific objectives can compete, making joint optimization challenging (Deng et al., 2025).

Recent advances in latent diffusion language models have shown promising results in generating text by denoising continuous representations (Dieleman, 2026). These models use a range of text representations, including compressed VAE latents (Guo et al., 2026; Jiang et al., 2026) and contextual T5 embeddings (Hu et al., 2026). Since latent diffusion is already widely used for visual generation (Rombach et al., 2022; Esser et al., 2024; Labs et al., 2025), these advances make it possible to model both modalities through continuous latent denoising. This motivates examining whether modeling both language and vision through continuous latent denoising can ease joint optimization in UMMs. We refer to this formulation as *unified latent diffusion*.

The benefits and trade-offs of unified latent diffusion for UMMs remain unclear. Although prior studies have demonstrated its feasibility (Bao et al., 2023; Xu et al., 2023; Yeo et al., 2026; Guo et al., 2026), controlled comparisons with UMMs using autoregressive language modeling remain limited. Their visual understanding evaluations also focus primarily on captioning, with broader capabilities such as VQA and OCR insufficiently explored. This leaves two questions open: Does unifying language and vision through latent diffusion ease multimodal training? What does this choice gain or lose in language generation and visual understanding?

In this work, we study unified latent diffusion through controlled multimodal pretraining experiments, training all model backbones from scratch. We compare autoregressive, discrete diffusion, and continuous latent diffusion language modeling using the same training data and matched model scales, while keeping visual modeling fixed. Building on the Transfusion (Zhou et al., 2025) framework, we use ELF (Hu et al., 2026) for latent diffusion language modeling (Fig. 1). We evaluate language generation, visual understanding, and visual generation to examine how the choice of language modeling affects joint training and task performance.

We share key findings:

- **Unifying generative modeling eases joint multimodal training and improves visual generation.** The choice of language modeling affects how well the model learns to generate images. With visual modeling held fixed, unified latent diffusion exceeds the autoregressive baseline's 30B-token DPG score after only 20B training tokens, suggesting that a common denoising formulation can benefit joint learning.
- **Iterative denoising strengthens visual understanding.** By progressively denoising text latents, the model can refine its answer. At the same 30B-token training budget, unified latent diffusion achieves a Vision-Centric score of 33.76, outperforming the autoregressive baseline's 20.84 by 12.92 points.
- **Language latent space matters, but not in the same way as in vision.** Latent selection substantially affects visual understanding, yet contextual T5 features offer little improvement over T5 vocabulary embeddings (25.3 vs. 24.8 on average). This contrasts with the gains from semantic representations reported in visual diffusion (Zheng et al., 2026), suggesting that the properties that make a latent space effective may differ across modalities.

2 Related Work

We review latent diffusion models, continuous diffusion language models, and unified multimodal models, focusing on representation spaces and generation mechanisms. A more extensive review is provided in Appendix A.

2.1 Latent Diffusion Models.

LDMs underpin many large-scale visual generators by modeling learned representations rather than pixels (Rombach et al., 2022; Esser et al., 2024). Beyond computational efficiency, the latent space shapes the distribution that the denoiser must learn, making representation design a key component of generative modeling. Representation autoencoders (Zheng et al., 2026) (RAEs) leverage

pretrained representation encoders to construct semantically rich latent spaces, showing that such representations can support faster convergence and higher-quality image generation. This motivates extending latent-space design to language generation in UMMs, where linguistic or vision-aligned representations can serve as denoising targets rather than merely input features.

2.2 Continuous diffusion language models.

Continuous latent diffusion models denoise Gaussian-corrupted text representations before decoding them into tokens (Lovelace et al., 2023). CoLa (Guo et al., 2026) uses Text VAE latents, while ELF applies flow matching to contextual embeddings (Hu et al., 2026). In the full-sequence formulation considered here, generation starts from Gaussian noise and jointly refines all target positions at each step. Bidirectional attention allows each position to incorporate evolving sequence context and available conditioning, enabling *sequence-wide latent refinement* with all positions remaining revisable until final token decoding. Flexible text representations and continuous diffusion sampling techniques, such as classifier-free guidance.

2.3 Unified Multimodal Models.

UMMs integrate visual understanding and generation within a single model. Existing approaches explore diverse visual generation paradigms, including autoregressive image-token prediction in Chameleon (Team, 2024), discrete diffusion in Show-o (Xie et al., 2025), and continuous diffusion or flow matching in Transfusion (Zhou et al., 2025) and BAGEL (Deng et al., 2025). Despite these differences, these models retain autoregressive text generation. Diffusion-based UMMs such as MMaDA (Yang et al., 2026a) extend language generation to masked diffusion. However, the language modeling design space remains comparatively underexplored, particularly continuous latent diffusion in UMMs. To our knowledge, no controlled study has compared autoregressive, discrete-diffusion, and continuous latent-diffusion language modeling in multimodal settings under matched model and data scales. This motivates a systematic study of language modeling formulations and latent representations for UMMs.

3 Unified Latent Diffusion Setup

Unified latent diffusion generates both vision and language through continuous latent diffusion. Following standard UMM settings, we consider three tasks (Fig. 2): (i) *language* (unconditional text generation or text-to-text), (ii) *visual understanding* (image-to-text, image+text-to-text) and (iii) *visual generation* (text-to-image).

Flow matching. Both modalities are trained with a flow matching objective (Lipman et al., 2022; Liu et al., 2022). Given a clean latent $\mathbf{x}$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, the noisy latent at time $t \in [0, 1]$ is $\mathbf{z}_t = t\,\mathbf{x} + (1-t)\,\boldsymbol{\epsilon}$, where $t{=}0$ is pure noise and $t{=}1$ is clean data. We use x-prediction: the model predicts the clean latent $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t)$, and the velocity is recovered as $\hat{\mathbf{v}} = (\hat{\mathbf{x}}_\theta - \mathbf{z}_t)/(1{-}t)$. This parameterization is more effective than velocity or noise prediction when the latent space is high-dimensional (Li & He, 2026; Tong et al., 2026a). The training objective is:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{x},\boldsymbol{\epsilon}} \left\| \hat{\mathbf{x}}_\theta(\mathbf{z}_t, t) - \mathbf{x} \right\|^2 . \tag{1}$$

Vision modeling. Following Tong et al. (2026a), we use a single RAE (Zheng et al., 2026) with a frozen SigLIP2 (Tschannen et al., 2025) vision encoder, for both visual generation and understanding. Images are encoded by the SigLIP2 encoder, and a trained decoder reconstructs images from these latents. For visual generation, these latents are generated with text conditioning. For visual understanding, the model generates text latents conditioned on these visual latents.

Language modeling. The language modeling component generates text for both *language* and *visual understanding* tasks; the latter additionally conditions on visual inputs. We adopt ELF (Hu et al., 2026), a flow matching based language model that has shown effective generation in high-dimensional latent spaces and supports denoising over long sequences. Its key techniques include self-conditioning (Chen et al., 2022; Team et al., 2026), training-time CFG (classifier-free guidance), and a shared-weight network that serves as both denoiser and decoder. See Appx. B for details. We

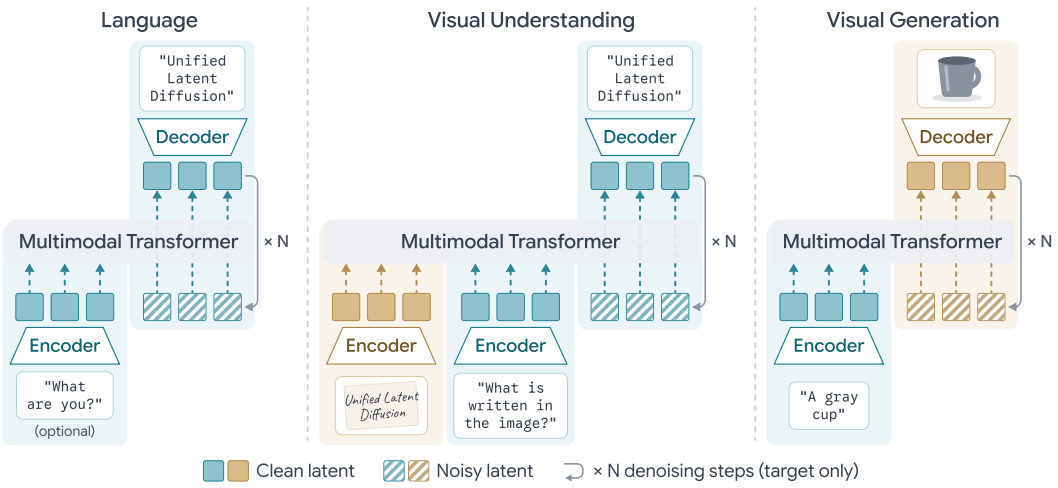

Figure 2: **Unified latent diffusion for three multimodal tasks.** A shared multimodal transformer iteratively denoises noisy latents into clean latents, with clean conditioning context. The same denoising mechanism is used for language generation (left), visual understanding (center), and visual generation (right), with modality-specific encoders and decoders.

use T5 (Raffel et al., 2020) for the latent encoder by default, based on the ablation experiments in Sec 5.3.

Training. Vision and language latent tokens are packed into a single sequence. We use block-wise causal attention (Zhou et al., 2025; Tong et al., 2026a): bidirectional within each modality segment, causal across segments. That is, the denoising segment attends to all preceding segments regardless of modality. Both modalities are trained with the flow matching objective (Eq. 1). We note that training-time CFG and latent decoding are applied for language loss (Appx. B). The overall loss combines per-modality losses:

$$\mathcal{L} = \lambda_v \, \mathcal{L}_{\text{vision}} + \lambda_l \, \mathcal{L}_{\text{language}}. \tag{2}$$

We use $\lambda_v = \lambda_l = 1.0$ by default.

4 Experiments

Our goal is to gain a deeper understanding of unified latent diffusion. Thus, we compare unified latent diffusion against UMMs with autoregressive and discrete diffusion language modeling. We fix vision modeling (Sec. 3) and vary only the language modeling method. We train all models from scratch with the same dense backbone, training data, training budget, and loss weights ($\lambda_v = \lambda_l = 1$). We also explore existing language latent spaces in Sec 5.3.

4.1 Experimental Setup

Model Architecture. We use a single transformer backbone for both language and visual generation. Following Bagel (Deng et al., 2025), we adopt the Qwen2.5-1.5B (Qwen et al., 2025) LLM architecture with 1.3B transformer backbone parameters. We adopt only the architecture; all models are trained from scratch. To study the effect of unification, we keep the backbone fully dense and shared across modalities, avoiding designs that introduce additional parameters to each modality (Mixture-of-Transformers; Liang et al., 2024)), or to individual experts (Mixture-of-Experts).

Optimization. We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1. We use constant learning rate of 1×10^{-4} after 1,000 steps of linear warmup. We apply exponential moving average (EMA) update with a decay of 0.9999. Notably, both the architecture and the optimization settings follow standard practice for autoregressive modeling.

Table 1: **Multimodal evaluation at 30B training tokens.** We compare three language modeling methods for UMMs: Latent Diffusion, Discrete Diffusion, Autoregressive. All other components, including vision modeling, training data, and backbone architecture are identical. See Sec. 4.1 for details. Best available results are in bold.

Generative Modeling		Language	Visual Understanding				Visual Generation			
Language	Vision	PPL ↓	Gen ↑	Know ↑	OCR ↑	Vision-Centric ↑	DPG ↑	Diff. Loss ↓	FID ↓	CLIPScore ↑
Autoregressive		**14.68**	23.58	15.92	**27.33**	20.84	65.16	0.7628	56.83	**23.31**
Discrete Diff.	Latent Diff.	30.63	**24.75**	**29.45**	22.73	29.34	59.19	0.7947	65.06	17.22
Latent Diff.		24.30	24.42	27.26	15.71	**33.76**	**74.00**	**0.7035**	**46.61**	15.40

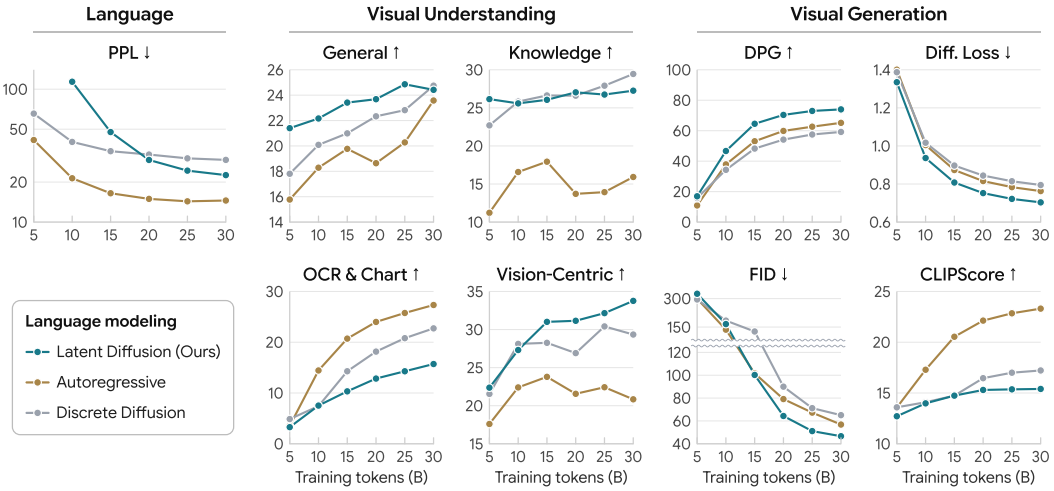

Figure 3: **Training dynamics of UMMs with different language modeling.** We compare three language modeling methods for UMMs: Latent Diffusion (Ours), Discrete Diffusion, Autoregressive. All other components, including vision modeling, training data, and backbone architecture are identical (Sec. 4.1). We track language, visual understanding, and visual generation performance throughout training.

Training Data. We use open-source data for all tasks. Training sets are sampled subsets from each source. Each batch consists of 0.5M tokens, sampled equally from the three sources. All images are processed at 224×224 resolution (256 tokens per image). (i) *language*: DataComp-LM (Li et al., 2024), 25.6M documents. Documents are chunked into 512-token sequences for training. For conditional generation, each text token sequence is split into a prefix condition and a denoising target, with sampled split ratio from $\mathcal{U}(0.1, 0.9)$. For unconditional generation, the entire sequence is a denoising target. (ii) *visual understanding*: DataComp-VLM (Farina et al., 2026), 56.2M samples, including both captioning and VQA data, which enables evaluation without additional fine-tuning. (iii) *visual generation*: a mixture of 10 datasets, 10.0M images (5 real-image, 3 synthetic, and 2 text-rendering datasets), with captions provided by i1 (Zeng et al., 2026). Following standard diffusion practice, we set the unconditional generation probability to 0.1 for all tasks. See Appx. C.2 for details.

Evaluation. We evaluate across three tasks. For unified latent diffusion, we use a 32-step Euler sampler and a CFG scale of 3.0 for both vision and language by default. (i) *language*: Perplexity (PPL) for generation. (ii) *visual understanding*: 16 benchmarks from the Cambrian benchmarks (Tong et al., 2024a), covering 4 categories: General, Knowledge, OCR & Chart, and Vision-Centric. (iii) *visual generation*: DPG-Bench (Hu et al., 2024), Diffusion Loss (flow matching loss) on 1K CC12M (Changpinyo et al., 2021) validation samples and FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021) on 30K COCO (Lin et al., 2014) validation samples.

4.2 Experiment Results

In Tab. 1, we evaluate UMMs with different language modeling on downstream benchmarks. We provide deeper analysis of unified latent diffusion in Sec. 5.

Language. Autoregressive modeling performs best on language generation, achieving the lowest generative perplexity of 14.68 (Tab. 1). Unified latent diffusion achieves a Gen. PPL of 24.30, outperforming discrete diffusion at 30.63, but remains behind autoregressive modeling.

Visual Understanding. Unified latent diffusion performs comparably to or better than autoregressive and discrete diffusion baselines on most visual understanding categories (Tab. 1). It achieves the highest Vision-Centric score (33.76), outperforming discrete diffusion (29.34) and autoregressive (20.84). On General and Knowledge benchmarks, unified latent diffusion and discrete diffusion perform similarly and both outperform autoregressive modeling. However, autoregressive modeling leads on OCR & Chart benchmarks (27.33 vs. 15.71).

Visual Generation. Unified latent diffusion substantially outperforms both baselines on visual generation. It achieves the best DPG score (74.00 vs. 65.16 for AR) and FID (46.61 vs. 56.83), showing stronger prompt following and higher image quality. Autoregressive modeling obtains the highest CLIPScore (23.31).

5 Analysis of Unified Latent Diffusion

5.1 Does Unifying Generative Objectives Ease Multimodal Training?

Does using the same generative formulation for vision and language make them easier to learn together? Under unified latent diffusion, the model learns the same type of probability path (Lipman et al., 2022; 2024) in both modalities: denoising trajectories from noise to clean latents, each in its own representation space. We examine whether this unification makes the two generative tasks easier to learn together.

Visual generation provides a useful setting for this comparison. Language generation and visual understanding both require producing text, so changing the language modeling method directly changes how these tasks are performed. In contrast, all three UMMs use the same visual representation, image generation objective, and sampling procedure. With the backbone architecture and training data also held fixed, visual generation lets us examine how the choice of language modeling affects the visual task learned alongside it.

Unified latent diffusion achieves lower validation visual diffusion loss than both baselines at every evaluated training budget (Fig. 3). Notably, at 20B training tokens, unified latent diffusion already surpasses the 30B-token autoregressive baseline in DPG score and achieves a lower visual diffusion loss, using one-third fewer training tokens. At 30B training tokens, it also achieves better FID and higher DPG scores (Tab. 1). These results show that adopting latent diffusion for language benefits visual learning within the shared model. The gains are not uniform across metrics: CLIPScore is lower, a trade-off we discuss in Sec. 5.2.

One interpretation is that unifying the generative tasks reduces *modality competition*—interference between learning signals (Deng et al., 2025). Prior works address this challenge through modality-specific parameterization (Deng et al., 2025; Fu et al., 2026b; Tong et al., 2026a). Our results suggest that unifying generative formulations may offer another way to improve joint training.

5.2 Latent Diffusion for Language Modeling: Gains and Losses

Sec. 5.1 has shown that unifying generative modeling into latent diffusion improves overall training dynamics of visual generation. However, a general-purpose UMM must also perform well on language and visual understanding tasks. We now examine what adopting latent diffusion for language modeling gains and loses relative to autoregressive modeling.

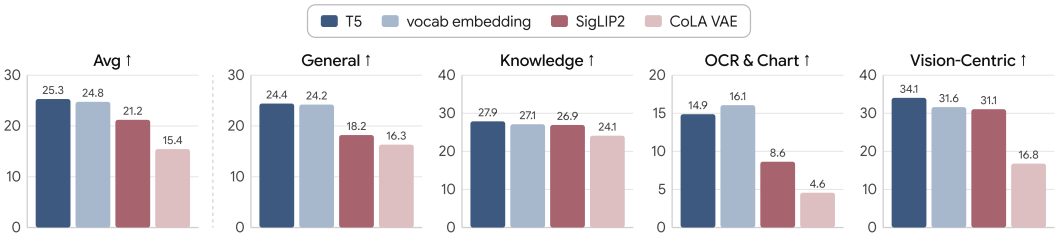

Figure 4: **Language latent space ablation on visual understanding.** We compare four language latent representations for latent diffusion language modeling: T5 encoder features, T5 vocabulary embeddings, CoLA VAE latents, and SigLIP2 text encoder features. All models are trained on language and visual understanding tasks only.

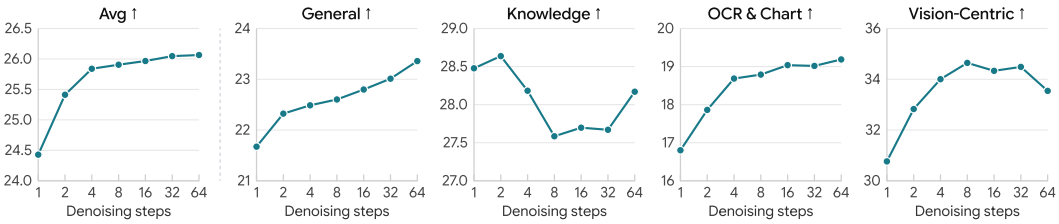

Figure 5: **Effect of denoising steps on visual understanding.** Visual understanding performance of unified latent diffusion as a function of the number of denoising steps. Increasing steps improves performance across categories, with particularly large gains on Vision-Centric benchmarks.

Latent diffusion falls behind in language generation. Prior studies find that autoregressive models require less training compute than continuous diffusion models to reach comparable likelihoods (Gulrajani & Hashimoto, 2023; Yang et al., 2026b). In our multimodal setting, autoregressive modeling also achieves better text generation performance at the same training-token budget (Tab. 1). Thus, the visual generation gains from unified latent diffusion come with a loss in text generation performance relative to autoregressive modeling.

Beyond language generation, the UMM with autoregressive language modeling also performs better on OCR & Chart in visual understanding (27.33 vs. 15.71) and on CLIPScore in visual generation (23.31 vs. 15.40). One possible explanation is that autoregressive modeling learns stronger language priors within the available training budget, which may also benefit these tasks.

Iterative denoising strengthens visual understanding. "*Visual understanding*" is about how the model can reason and think correctly about the given visual observation. Latent diffusion offers an advantage here: it refines all tokens over multiple denoising steps, giving the model repeated opportunities to revisit and correct its reasoning. Fig. 6 illustrates the role of this iterative process: the model gradually narrows its attention to the region that contains the answer. This refinement enables test-time scaling: increasing the number of denoising steps improves overall visual understanding, with particularly large gains on Vision-Centric benchmarks (Fig. 5). These benchmarks rely more on spatial reasoning than language knowledge, which is where iterative refinement compensates for latent diffusion's weaker language capability.

Parallel denoising enables length-controlled generation. Latent diffusion denoises all target tokens in parallel, so the number of tokens to generate is fixed before generation begins. Autoregressive modeling has no such control: next token generation simply continues until an end-of-sequence (EOS) token appears. A target length can be requested through instructions, but this is an indirect, data-driven solution rather than a property of the model. Fig. 7 compares the two. We note that this control does not sacrifice flexibility: latent diffusion can also be trained to emit EOS tokens, so both explicit control and model-decided length are available. We note that latent diffusion supports both modes: it can also be trained to emit EOS tokens, letting the model decide the length itself (Guo et al., 2026).



Figure 6: **Attention map over denoising steps on VQA.** We visualize the attention from the denoising text token to image tokens. As denoising progresses, the model gradually focuses attention on the image region relevant to the answer.

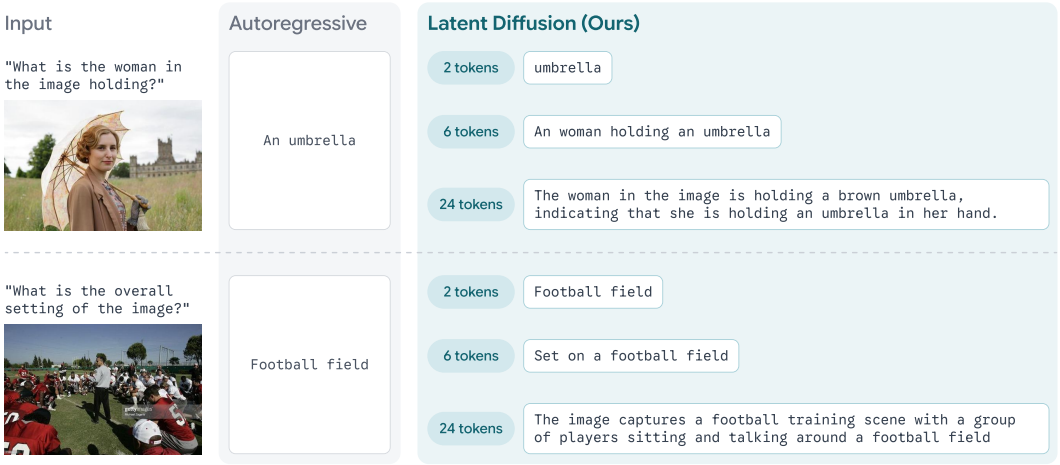

Figure 7: **Length-controlled generation in VQA.** Autoregressive modeling generates tokens until EOS, offering no direct control over output length. Latent diffusion generates a fixed number of tokens in parallel, enabling explicit length control (shown with 2, 6, and 24 tokens).

5.3 Exploring Language Latent Spaces

In visual generation, the choice of latent space substantially affects diffusion model quality (Zheng et al., 2026). For language, the role of latent representations in continuous diffusion remains under-explored, especially in multimodal settings. We therefore compare language latent spaces, examining whether contextual semantic features offer benefits over vocabulary embeddings.

Experiment Setup. To focus on language capabilities and reduce experimental cost, we train on language and visual understanding tasks only, excluding visual generation. Other settings follow Sec. 4.1. We compare four representations: T5 vocabulary embeddings (Raffel et al., 2020), compressed CoLA VAE latents (Guo et al., 2026), contextual T5 encoder features, and vision-aligned SigLIP2 text encoder features (Tschannen et al., 2025). We find that reconstructing text from SigLIP2's features is difficult (approximately 90% reconstruction accuracy, compared to 99% for

the other encoders). Inspired by RAEv2 (Singh et al., 2026), we define the SigLIP2 latent as the average of its vocabulary embeddings and final-layer features, to improve reconstruction accuracy.

Results. Fig. 4 shows that latent space selection affects visual understanding performance: CoLA VAE and SigLIP2-based latents perform notably worse than T5-based representations. However, T5 final-layer features and T5 vocabulary embeddings achieve similar average visual understanding performance. Unlike the substantial gains reported for semantic representations in visual diffusion (Zheng et al., 2026), contextual semantic features provide limited additional benefit in our language diffusion setting.

One possible explanation is that language tokens already represent abstract symbolic information, whereas images consist of low-level visual signals. Mapping tokens into contextual semantic representations may therefore offer smaller additional benefits. Under our ELF-based setup, these findings suggest that contextual semantics and vision alignment alone do not determine the suitability of a language latent space. Understanding which representation properties benefit language diffusion, and how they interact with denoising and decoding, remains an important direction for future work.

6 Limitations and Future Work

Training scale and data diversity. Our experiments are limited to a single model scale and a fixed training setup. Exhaustively evaluating different model sizes, longer training budgets, and substantially larger or more diverse data mixtures would require considerable compute, especially because each setting must be retrained across multiple language modeling paradigms for a controlled comparison. It therefore remains unclear whether the observed trends persist at substantially larger scales or under different data compositions. Extending these comparisons to larger models, longer training, and broader multimodal data mixtures is an important direction for future work.

Maturity of continuous language modeling. Continuous language modeling is still developing rapidly, and its current architectures, objectives, and decoding procedures may be far less optimized than those of autoregressive language models, which have benefited from decades of refinement. The performance gaps we observe in Generative perplexity, OCR & Chart, and CLIPScore may therefore partly reflect limitations of the current continuous modeling design rather than an inherent limitation of the paradigm itself. As continuous language models continue to improve, revisiting these comparisons with stronger architectures and training objectives will be important.

7 Conclusion

We study Unified Latent Diffusion for multimodal modeling as a generative formulation for UMMs. Through controlled comparisons with autoregressive and discrete diffusion language modeling, we find that this unification changes the trade-offs of multimodal learning. Unified latent diffusion substantially improves visual generation and achieves strong visual understanding, while iterative denoising further enables test-time refinement. Our results suggest that language latent spaces may benefit less from semantic representations than their visual counterparts. Overall, latent diffusion may offer a useful direction for unified multimodal learning, while providing capabilities beyond conventional autoregressive generation.

Reproducibility statement

We provide detailed descriptions of model architectures in Appendix C.1, training configurations including datasets, preprocessing, and hyperparameters in Appendix C.2, and evaluation protocols and benchmarks in Appendix C.3. We will release the full codebase and pretrained checkpoints upon acceptance.

AI use statement

In this work, we used generative AI tools for writing assistance, including grammar checking, improving clarity, and polishing author-written text. We also used these tools to review and debug

code used in experiments. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims, and experimental results.

REFERENCES

Inclusion AI, Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, et al. Llada2. 0-uni: Unifying multimodal understanding and generation with diffusion large language model. *arXiv preprint arXiv:2604.20796*, 2026.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

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.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. 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, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Sander Dieleman. Continuous diffusion language models, 2026. URL `https://sander.ai/2026/08/24/continuous-dlms.html`.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

Matteo Farina, Vishaal Udandarao, Thao Nguyen, Selim Kuzucu, Maximilian Böther, Andreas Hochlehnert, Adhiraj Ghosh, Marianna Nezhurina, Karsten Roth, Joschka Struber, et al. Datacomp-vlm: Improved open datasets for vision-language models. *arXiv preprint arXiv:2606.28551*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, et al. Lance: Unified multimodal modeling by multi-task synergy. *arXiv preprint arXiv:2605.18678*, 2026b.

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.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16693–16715. Curran Associates, Inc., 2023. doi: 10.52202/075280-0730. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/35b5c175e139bff5f22a5361270fce87-Paper-Conference.pdf.

Hongcan Guo, Qinyu Zhao, Yian Zhao, Shen Nie, Rui Zhu, Qiushan Guo, Feng Wang, Tao Yang, Hengshuang Zhao, Guoqiang Wei, et al. Continuous latent diffusion language model. *arXiv preprint arXiv:2605.06548*, 2026.

Junlin Han, Shengbang Tong, David Fan, Minghao Chen, Philip Torr, Filippos Kokkinos, and Mike Lewis. Towards physics of multimodal pretraining: Knowledge flow, modality synergy, early unification, and recipes. *arXiv preprint arXiv:2608.05000*, 2026.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11575–11596, 2023.

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.

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. *Advances in neural information processing systems*, 30, 2017.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

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.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Jiaxiu Jiang, Jingjing Ren, Wenbo Li, Bo Wang, Haoze Sun, Yijun Yang, Jianhui Liu, Yanbing Zhang, Shenghe Zheng, Yuan Zhang, et al. Textldm: Language modeling with continuous latent diffusion. *arXiv preprint arXiv:2605.07748*, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Jaeik Kim, Woojin Kim, Jihwan Hong, Yejoon Lee, Sieun Hyeon, Mintaek Lim, Yunseok Han, Dogeun Kim, Hoeun Lee, Hyunggeun Kim, et al. Dynin-omni: Omnimodal unified large diffusion language model. *arXiv preprint arXiv:2604.00007*, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

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.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13178–13188. IEEE, 2025a.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2779–2790. IEEE, 2025b.

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.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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 *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15740–15751, 2026.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36:56998–57025, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26429–26445. IEEE, 2024a.

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 *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. Tess: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2347–2361, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: a novel resource for question answering on scholarly articles: T. saikh et al. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

Ying Shen, Tianrong Chen, Yuan Gao, Yizhe Zhang, Yuyang Wang, Miguel Ángel Bautista, Shuangfei Zhai, Joshua M Susskind, and Jiatao Gu. Starflow2: Bridging language models and normalizing flows for unified multimodal generation. *arXiv preprint arXiv:2605.08029*, 2026.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 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.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. Self-conditioned embedding diffusion for text generation, 2022. URL https://arxiv.org/abs/2211.04236.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

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

DiffusionGemma Team, Adrien Ali Taïga, James Assiene, Daniele Calandriello, Rahma Chaabouni, João Gante, Tamara von Glehn, Nate Keating, Chris Knutsen, Martin Kukla, et al. Diffusiongemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024b.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025.

Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 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 arXiv:2601.16208*, 2026b.

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.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *Advances in Neural Information Processing Systems*, 38:81402–81440, 2026.

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*, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv e-prints*, pp. arXiv–2506, 2025b.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

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. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025c.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

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.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7754–7765, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026a.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. *URL https://arxiv. org/abs/2508.15487*.

Kyeongmin Yeo, Yunhong Min, and Minhyuk Sung. Muni: Multimodal unified latent diffusion for coherent any-to-any generation. *arXiv preprint arXiv:2606.16408*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Boya Zeng, Tianze Luo, Shu Pu, Jucheng Shen, Taiming Lu, Gabriel Sarch, and Zhuang Liu. i1: A simple and fully open recipe for strong text-to-image models. *arXiv preprint arXiv:2606.11289*, 2026.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. Planner: Generating diversified paragraph via latent language diffusion model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 80178–80190. Curran Associates, Inc., 2023. doi: 10.52202/075280-3513. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fdba5e0a9b57fce03e89cc0cad0a24e9-Paper-Conference.pdf.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

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*, volume 2025, pp. 6446–6469, 2025.

A Extensive Related Work

A.1 Unified Multimodal Models

We organize UMMs by their language and image generation formulations, distinguishing the generative backbone from modality-specific encoders and output decoders.

Autoregressive language and image generation. Chameleon (Team, 2024), Emu3 (Wang et al., 2024), and VILA-U (Wu et al., 2025c) model text and discrete visual tokens through autoregressive prediction. Janus (Wu et al., 2025a) and Janus-Pro (Chen et al., 2025b) retain this formulation while separating visual encoders for understanding and generation. Unified-IO 2 (Lu et al., 2024a) extends autoregressive modeling to additional modalities, including audio and actions. Autoregressive visual generation also accommodates continuous outputs: UniFluid (Fan et al., 2025) generates continuous image tokens using a per-token diffusion head, while STARFlow2 (Shen et al., 2026) uses autoregressive normalizing flows. These methods retain autoregressive language modeling despite differing visual representations and prediction objectives.

Autoregressive language with discrete image diffusion. Show-o (Xie et al., 2025) combines autoregressive text prediction with masked diffusion over image tokens in a shared transformer. Emu3.5 (Cui et al., 2025) similarly preserves autoregressive language generation while adapting an autoregressively pretrained visual generator to discrete diffusion. Both replace sequential image-token prediction with parallel denoising without changing the language generation formulation.

Autoregressive language with continuous image diffusion. Transfusion (Zhou et al., 2025), JanusFlow (Ma et al., 2025), Show-o2 (Xie et al., 2026), and BAGEL (Deng et al., 2025) combine autoregressive text generation with continuous visual diffusion or flow matching. Their visual interfaces and parameter-sharing strategies differ. BAGEL separates semantic understanding features from VAE generation latents and uses specialized transformer experts. Mixture-of-Transformers (Liang et al., 2024) likewise introduces modality-specific parameters while preserving cross-modal attention. Tuna (Liu et al., 2026) connects VAE and representation encoders for shared visual processing while generating VAE latents. RAE-based systems (Tong et al., 2026b;a) instead use pretrained semantic features as both understanding inputs and generation targets. SenseNova-U1 (Diao et al., 2026) performs visual flow matching directly in pixel space. These approaches extensively vary visual modeling while retaining autoregressive language generation.

Autoregressive language with external diffusion generators. Other systems connect language-based understanding to separate visual generators rather than performing both tasks within one denoising backbone. SEED-X (Ge et al., 2024) and MetaMorph (Tong et al., 2025) predict visual embeddings that condition diffusion decoders. BLIP3-o (Chen et al., 2025a) learns a flow-matching prior over CLIP image embeddings conditioned on a vision–language backbone, followed by image decoding. OmniGen2 (Wu et al., 2025b) conditions a separate image diffusion transformer on language-model hidden states. NExT-GPT (Wu et al., 2023) extends this modular approach to image, audio, and video outputs through modality-specific diffusion decoders.

Discrete-state generation for language and images. UniDisc (Swerdlow et al., 2025), MMaDA (Yang et al., 2026a), Muddit (Shi et al., 2025), and Lumina-DiMOO (Xin et al., 2025) jointly model discrete text and image tokens through diffusion, supporting image understanding and synthesis alongside capabilities such as infilling and editing. LLaDA2.0-Uni (AI et al., 2026) uses block-level masked diffusion over text and semantic visual tokens, with a separate diffusion decoder reconstructing images. FUDOKI (Wang et al., 2026) instead uses discrete flow matching for both outputs while retaining separate visual representations for understanding and generation. Dynin-Omni (Kim et al., 2026) extends discrete diffusion to text, image, and speech understanding and generation, additionally supporting video understanding.

Discrete language diffusion with continuous image diffusion. Dual Diffusion (Li et al., 2025b) combines masked diffusion for language with continuous image denoising trained through flow matching. Although both modalities use denoising-based generation, their state spaces and pre-

diction targets remain distinct: categorical tokens for language and continuous representations for images.

Continuous generation of language and image representations. UniDiffuser (Bao et al., 2023) jointly models continuous image and text representations using modality-specific diffusion times. OmniFlow (Li et al., 2025a) extends rectified flow to text, images, and audio, using a Text VAE with an autoregressive text decoder. CoLa (Guo et al., 2026) presents a preliminary text–image prototype with modality-specific VAEs and a shared block-causal continuous latent prior. These establish precedents for continuous multimodal generation, although their visual understanding demonstrations center on captioning or qualitative examples rather than broad instruction-following VQA. Continuous latent generation should also be distinguished from the mechanism used to decode latents into text.

A.2 Representation Autoencoders and Latent-Space Design

Representation alignment and representation autoencoders. Visual diffusion research increasingly treats representation design as a component of generative modeling, rather than merely an interface for compressing images. Pretrained visual representations have been incorporated both as supervision for the denoiser and as the space in which generation takes place. REPA (Yu et al., 2024) follows the former approach, aligning intermediate denoiser features computed from noisy inputs with clean-image representations from pretrained visual encoders. This auxiliary supervision improves training efficiency and generation quality without replacing the underlying latent targets. REPA-E (Leng et al., 2025) extends representation alignment to joint tokenizer–denoiser adaptation, blocking diffusion-loss gradients at the tokenizer to prevent the latent degradation that naive end-to-end optimization can cause.

Representation autoencoders. A complementary line of work uses pretrained representations as the generative targets themselves, replacing the conventional VAE latent space. RAEs (Zheng et al., 2026) pair a frozen representation encoder, such as DINO, SigLIP, or MAE, with a trainable reconstruction decoder and learn diffusion directly over the resulting semantic features. This changes not only the supervision provided to the denoiser, but also the distribution it must generate. The original RAE study shows that semantically rich representations can support faithful reconstruction and strong generation, provided that denoiser capacity, noise scheduling, and decoder robustness are adapted to the high-dimensional latent space. RAEv2 (Singh et al., 2026) further improves representation extraction by aggregating multiple late encoder layers without increasing the latent tensor dimensions. It also demonstrates complementary benefits from REPA, indicating that choosing informative generative targets and supervising the denoiser's internal representations address distinct aspects of generative learning.

Extending this approach to large-scale text-to-image generation, Tong et al. (2026b) report faster convergence and improved generation quality with RAE latents relative to VAE-based baselines. Their framework also shares visual representations between understanding and generation, while retaining autoregressive language modeling. Together, these studies establish representation space as an important design choice for visual generation, motivating a corresponding investigation of language generation in UMMs.

A.3 Diffusion Language Models

Discrete Diffusion Language Models D3PM (Austin et al., 2021) formulates categorical diffusion with absorbing and uniform corruption. Subsequent work develops continuous-time formulations (Campbell et al., 2022). MDLM (Sahoo et al., 2024) and Shi et al. (2024) simplify absorbing-state diffusion into weighted masked-token prediction. Iteratively unmasking subsets of positions enables *any-order and parallel generation*, leveraging bidirectional context while potentially reducing sequential model evaluations.

LLaDA (Nie et al., 2026) establishes from-scratch masked-diffusion pretraining and instruction tuning at the 8B scale. Further scaling includes autoregressive-to-diffusion adaptation in Dream 7B (Ye et al.), high-speed code generation in Seed Diffusion (Song et al., 2025), and models with up to 100B

total parameters in LLaDA 2.0 (Bie et al., 2025). DiffusionGemma (Team et al., 2026) combines denoising fine-tuning, reinforcement learning, and sampler distillation for parallel block generation.

Continuous Diffusion Language Models Continuous diffusion language models differ in their representation spaces, training objectives, and decoding mechanisms. Diffusion-LM (Li et al., 2022) and DiffuSeq (Gong et al., 2022) denoise token embeddings for controllable and sequence-to-sequence generation, respectively. Self-conditioned Embedding Diffusion (Strudel et al., 2022) reuses previous predictions during denoising, while CDCD (Dieleman et al., 2022) trains continuous-state diffusion with token-level cross-entropy. SSD-LM (Han et al., 2023) and TESS (Mahabadi et al., 2024) instead use vocabulary-aligned simplex representations, with block-wise and full-sequence generation, respectively.

Latent-space approaches separate representation learning from generation. LD4LG (Lovelace et al., 2023) uses pretrained encoder–decoder representations, while PLANNER (Zhang et al., 2023) combines latent diffusion planning with autoregressive decoding. CoLa (Guo et al., 2026) uses Text VAE latents with a block-causal flow-matching prior and conditional decoding. ELF (Hu et al., 2026) instead applies flow matching to contextual embeddings, using clean-embedding prediction and shared-weight token decoding. In full-sequence formulations such as ELF, generation jointly updates all target positions from Gaussian noise, enabling *sequence-wide latent refinement* through bidirectional interactions until final token decoding. Together with sampling controls such as classifier-free guidance (Dieleman et al., 2022; Hu et al., 2026), these approaches make representation selection and iterative refinement explicit design choices for language generation.

B Language modeling: Embedded Language Flows

This appendix summarizes the ELF-based language modeling component of our framework (Hu et al., 2026), focusing on shared-weight denoising, decoding, and self-conditioning.

B.1 ELF Framework

Shared-weight denoising and decoding. Given a token sequence $\mathbf{s}$ and its encoder latents $\mathbf{x} = E(\mathbf{s})$, the forward process constructs noisy latents via linear interpolation: $\mathbf{z}_t = t\,\mathbf{x} + (1-t)\,\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \in [0,1]$. The denoiser is trained with the x-prediction objective $\mathcal{L}_{\text{flow}}$ (Eq. 1). Predicting clean latents directly is well-suited to high-dimensional representation spaces and enables weight sharing between the denoiser and the token decoder (Hu et al., 2026).

Concretely, the decoder reuses the denoising Transformer backbone, augmented with a vocabulary projection head that produces token logits $\boldsymbol{\ell}_\theta^{\text{dec}}$. To ensure robustness to imperfect latents at generation time, the decoder is trained on token-wise corrupted latents $\tilde{\mathbf{z}}$:

$$\mathcal{L}_{\text{CE}} \;=\; \mathbb{E}_{\mathbf{s},\,\tilde{\mathbf{z}}}\big[\,\text{CE}\big(\boldsymbol{\ell}_\theta^{\text{dec}}(\tilde{\mathbf{z}}),\;\mathbf{s}\big)\big]\,, \tag{3}$$

where the cross-entropy loss is averaged over all target token positions. During decoding, the model operates at $t{=}1$ without self-conditioning.

Control tokens and sampling. The diffusion timestep t, the self-conditioning guidance scale ω_{sc}, and a binary mode flag $m \in \{\texttt{denoise}, \texttt{decode}\}$ are encoded as three control tokens and prepended to the input sequence. At inference, sampling begins from Gaussian noise and iteratively refines the latents using the velocity estimate $\hat{\mathbf{v}} = (\hat{\mathbf{x}}_\theta - \mathbf{z}_t)\,/\,(1-t)$, while external conditioning context remains fixed throughout. Only the final denoised latent is passed to the decoder, which produces the output token sequence via position-wise argmax.

B.2 Classifier-Free Guidance

Self-conditioning and training-time CFG. Self-conditioning (Chen et al., 2022) concatenates a clean-latent estimate $\mathbf{h}$ with the noisy latent $\mathbf{z}_t$ along the feature dimension, followed by a learned linear projection. During training, $\mathbf{h}$ is stochastically set to either $\text{sg}(\hat{\mathbf{x}}_\theta^0)$, a preliminary prediction without self-conditioning, or $\mathbf{0}$, where sg denotes stop-gradient. Let $\hat{\mathbf{x}}_\theta^{\text{sc}}$ denote the subsequent self-conditioned prediction. For a sampled guidance scale $\omega_{\text{sc}} > 0$, ELF replaces the clean target in the

denoising objective with

$$\mathbf{x}_{\text{tgt}} \;=\; \text{sg}\Big[\,\mathbf{x} + \Big(1 - \tfrac{1}{\omega_{\text{sc}}}\Big)\big(\hat{\mathbf{x}}_\theta^{\text{sc}} - \hat{\mathbf{x}}_\theta^{0}\big)\Big]\,. \tag{4}$$

Setting $\omega_{\text{sc}}{=}1$ recovers standard self-conditioned training. Since guidance is learned during training, inference requires only a single denoiser evaluation per Euler step, where $\mathbf{h}$ is initialized to zero and reuses the previous step's prediction thereafter.

External-context CFG. External context provides a complementary guidance signal. Let $\hat{\mathbf{x}}_\theta^{\text{cond}}$ and $\hat{\mathbf{x}}_\theta^{\text{drop}}$ denote predictions with full and selectively dropped context, respectively. Context-based CFG computes

$$\hat{\mathbf{x}}_\theta^{\text{ctx}} \;=\; \hat{\mathbf{x}}_\theta^{\text{drop}} + \omega_{\text{ctx}}\Big(\hat{\mathbf{x}}_\theta^{\text{cond}} - \hat{\mathbf{x}}_\theta^{\text{drop}}\Big) \tag{5}$$

and substitutes this into the velocity computation, requiring two forward passes per step. In our multimodal setting, context $\mathbf{c}{=}(\mathbf{c}_{\text{v}}, \mathbf{c}_{\text{l}})$ comprises visual and language components. Dropping one component while retaining the other enables modality-specific guidance.

C Experimental Details

C.1 Architecture Details

Shared backbone. All models use the Qwen2.5-1.5B architecture (Qwen et al., 2025): 28 layers, hidden width $H = 1536$, FFN width 8960 with SiLU, and GQA with 12 query and 2 key–value heads. The backbone is trained from scratch with $\mathcal{N}(0, 0.02^2)$ weight initialization, zero biases, and untied input/output embeddings. It is fully shared across modalities, without modality-specific transformers or expert routing. Vision paths, image flow-matching heads, data, and optimization are identical across variants; only text modeling differs.

Vision paths. Both paths use a frozen SigLIP 2 So400m encoder (Tschannen et al., 2025) at 224×224 (27 blocks and the final LayerNorm, without attention pooling), producing 256 tokens of dimension 1152. For understanding, a two-layer MLP projects these features to the backbone width. For generation, the features serve as RAE latents without additional normalization or rescaling. Both paths add fixed 2D sin–cos embeddings after input projection; each image occupies a single RoPE position shared by its 256 tokens.

Text modeling. AR and MDLM use the Qwen2.5 tokenizer and token embeddings. AR minimizes next-token cross-entropy, whereas MDLM applies absorbing-state masking, leaving condition text unmasked. MDLM receives no timestep input and uses the SUBS parameterization. ELF uses the T5-base tokenizer and independently encodes each span with a frozen T5-base encoder into 768-dimensional latents. The concatenation $[z_t, \hat{x}_0^{\text{sc}}]$ is projected through three linear layers. The denoising head comprises RMSNorm and a zero-initialized linear layer ($H \to 768$). Each ELF text slot additionally receives $\text{Linear}_{3H\to H}([e_t, e_w, e_m])$, without adding sequence positions: e_t is the mean of four learned vectors plus a timestep embedding, e_w encodes the self-conditioning CFG scale, and e_m identifies conditions.

C.2 Training Data

We use three publicly available data sources for visual generation, visual understanding, and language modeling. All model variants use the same selected source data, with model-specific text representations and training objectives.

Visual generation: i1. We construct a 10M-image training set using the image sources and synthetic captions released by i1 (Zeng et al., 2026). Specifically, we select 1M images from each of ten sources: five real-image datasets (ImageNet-22K, YFCC100M, RedCaps, Places365-Challenge 2016, and Pexels), three synthetic-image datasets (FLUX-Reason-6M, GPT-Image-Edit-1.5M, and Midjourney v6), and two text-rendering datasets (RenderedText and TextAtlas). This yields an equal-sized subset from each source, stored in 12,800 shards.

Visual understanding: DCVLM. We use a training snapshot of the balanced DataComp-VLM collection (Farina et al., 2026), containing 5,656 shards and approximately 56.2M samples, with a nominal size of 100B multimodal tokens. DataComp-VLM aggregates 160 source datasets spanning image-caption pairs, multimodal documents, text-only data, and multimodal instruction data. The upstream balanced mixture assigns 40% of training samples to image captioning, 40% to visual instruction, 15% to text-only instruction, and 5% to interleaved multimodal documents.

Language modeling: DCLM. We use DCLM-Baseline (Li et al., 2024), a quality-filtered English web-text corpus derived from Common Crawl. Its upstream construction combines text extraction, heuristic cleaning, near-duplicate removal, and fastText-based quality filtering, with OpenHermes 2.5 and ELI5 examples used as positive references for the quality classifier. Our training set consists of 25 of the 100 shards in our `DCLM-baseline-100BT` snapshot, comprising approximately 25.6M documents. The corresponding corpus size is approximately 25B tokens, estimated from text bytes rather than an exact token count under each model's tokenizer.

C.3 Evaluation

Visual understanding. We evaluate the pretrained checkpoints directly, without instruction tuning, on 16 benchmark configurations using the prompts and rule-based scorers of VLMEvalKit (Duan et al., 2024) (v0.2rc1). Following Han et al. (2026), the benchmarks are grouped into *General* (GQA (Hudson & Manning, 2019), MME (Fu et al., 2026a), MMBench (Liu et al., 2024a), SEED-Bench (Li et al., 2023)), *Knowledge* (ScienceQA (Saikh et al., 2022), MMMU (Yue et al., 2024), AI2D (Kembhavi et al., 2016), MathVista (Lu et al., 2024b)), *OCR & Chart VQA* (TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), OCRBench (Liu et al., 2024b)), and *Vision-Centric* (RealWorldQA, MMVP (Tong et al., 2024b), CV-Bench 2D/3D (Tong et al., 2024a)).

Images are bicubically resized with preserved aspect ratio to a 224-pixel long side and centered on a black 224×224 canvas, matching the training transform; interleaved images retain their order. Prompts use VLMEvalKit's templates unchanged, with no system prompt or role tags: multiple-choice questions list the options and ask for the correct one, open-ended VQA appends "Answer the question using a single word or phrase.", and yes/no questions are passed as is. All scoring is local and rule-based (no LLM judge). Decoding follows each model's training objective. AR decodes greedily with 32 tokens. MDLM fills an answer slot of L mask tokens, unmasking one token per step (L steps) at the most confident position without re-masking. ELF denoises the answer latents with a 64-step stochastic sampler, classifier-free guidance of scale 3.0 and self-conditioning classifier-free guidance 1.0, and decodes the final latents with its decoder head.

Language evaluation. We evaluate language generation with prefix-conditioned generative perplexity on 100 held-out DCLM documents. They are drawn from a shard of DCLM-100BT (Li et al., 2024) that is not part of the 25B-token training subset, keeping the first 100 documents (in file order) that contain at least 512 tokens under both the Qwen2.5 and T5 tokenizers. The conditioning protocol is identical to the prefix conditioning used for DCLM during training. Each document contributes one 512-token chunk taken from its start, tokenized with the model's own tokenizer (Qwen2.5 for AR and MDLM, T5 for ELF), matching the training chunk length and document-chunk packing. A split fraction $f \sim \mathcal{U}(0.1, 0.9)$, the training range for the condition fraction, is drawn once per document and shared by all models. AR and MDLM sample with temperature 1 and top-$k = 50$. ELF uses a 64-step stochastic sampler with self-conditioning guidance 3.0 and no prefix guidance, followed by argmax decoding. Generations are scored by GPT-2 Large (Radford et al., 2019).

Text-to-image generation. We evaluate text-to-image generation using DPG, Diffusion Loss, FID, and CLIPScore. All models generate 224×224 images by integrating the flow from Gaussian noise with a 32-step Euler solver on a uniform time grid, without an inference-time timestep shift; the predicted clean latent is converted to a velocity with the denominator $1 - t$ clamped at 0.05. We apply classifier-free guidance with scale 3.0, where the unconditional input contains only the image latents, as in a caption-dropped training sample. Captions are formatted in each model's training layout, and the generated RAE latents are decoded by the frozen RAE decoder. Initial noise is seeded from a global seed of 42, so all models are evaluated from identical noise.

DPG-Bench (Hu et al., 2024) measures compositional prompt following over its 1,065 dense prompts (14,392 decomposed questions) using the i1 evaluation suite (Zeng et al., 2026): four images are generated per prompt, each question is answered by the mPLUG-large VQA model (questions whose parent question fails score 0), and the score averages over questions, images, and prompts. Diffusion Loss is the flow-matching training objective, computed without guidance on 1,000 image–text pairs randomly sampled from CC12M (Changpinyo et al., 2021) with a fixed seed 42. Images are center-cropped and resized to 224 pixels as in training, and each pair is evaluated at 10 timesteps $u_k = (k + 0.5)/10$, $k = 0, \ldots, 9$, mapped through the training timestep shift, with fixed noise per (pair, timestep) shared by all models. We measure distributional image quality and image–text alignment on COCO-30K (Lin et al., 2014): the first 30,000 COCO 2014 validation image IDs (sorted), each with one caption (the lowest annotation ID) and one generated image. FID is computed with `torch-fidelity` (Inception-v3 pool features) against the corresponding real images, center-cropped to square and Lanczos-resized to 224×224. CLIPScore is evaluated with OpenCLIP ViT-L/14.