MULTIMODAL FLOW: UNIFIED FLOW MODELING OF LANGUAGE AND VISION IN EMBEDDING SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **Multimodal Flow**, a fully continuous generative model of language and vision. Most unified multimodal models either model both language and quantized images as discrete tokens or combine discrete language prediction with continuous image generation. The former introduces a visual quantization bottleneck. The latter requires modality-dependent objectives and sampling procedures. Fully continuous modeling avoids these trade-offs and enables a shared generative process, but remains underexplored for multimodal pretraining. Multimodal Flow introduces a unified continuous architecture that integrates multimodal continuous representations with a shared chunk-causal flow backbone. It organizes text blocks and images as ordered continuous hyperchunks, preserving textual token order and visual spatial structure. The backbone learns a single vector field over these hyperchunks through Flow Matching. Joint attention enables cross-modal interaction, while modality-specific feed-forward networks process each modality. The model predicts multiple target chunks in parallel during training and generates hyperchunks sequentially at inference. We instantiate **MF-1** and pretrain it on multimodal data. Across 0.6B, 1.2B, and 1.6B scales, continued pretraining consistently improves multimodal modeling. With only 150B pretraining tokens, MF-1 achieves an average score of 82.8 across GenEval and DPG-Bench and 75.3 across VQAv2, MMBench, and POPE, remaining competitive with unified models trained on substantially more data. Under matched data, optimization, and parameter budgets, Multimodal Flow further outperforms representative hybrid and discrete models. These results establish continuous chunk-based embedding flow modeling as a new fully continuous paradigm for unified multimodal modeling. The code and model will be released to facilitate future research.

1 INTRODUCTION

Vision-language models have advanced visual understanding by conditioning text generation on images (Liu et al., 2023; Tao et al., 2025; Zeng et al., 2026). Recent unified multimodal models extend this setting by treating both text and images as generation targets within a single pretrained model (Team, 2024; Zhou et al., 2025; Wang et al., 2024b; Zou et al., 2025). These advances raise a broader question: *can language and vision share a generative process while preserving the representational fidelity and structure each requires?* Doing so is nontrivial because language is organized as an ordered sequence of tokens, whereas images have dense spatial structure, and the two have traditionally relied on different representations and generation mechanisms.

Current unified models resolve this tension through two dominant paradigms. Fully discrete models quantize images into visual tokens, placing both modalities in a shared categorical sequence (Team, 2024; Wang et al., 2024b; Xie et al., 2025). This alignment facilitates joint modeling, but makes visual fidelity dependent on the tokenizer: fine-grained details discarded by the tokenizer are unavailable to the model (Tang et al., 2026). Hybrid discrete–continuous models instead retain continuous visual states and combine discrete language prediction with diffusion or flow for images (Zhou et al., 2025; Ma et al., 2025; Li et al., 2025b). They avoid visual quantization and permit interaction within a shared backbone, but language and vision remain governed by different objectives and sampling procedures. As summarized in Figure 1, neither paradigm simultaneously provides continuous visual states and a common generative process across modalities. This motivates fully

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.

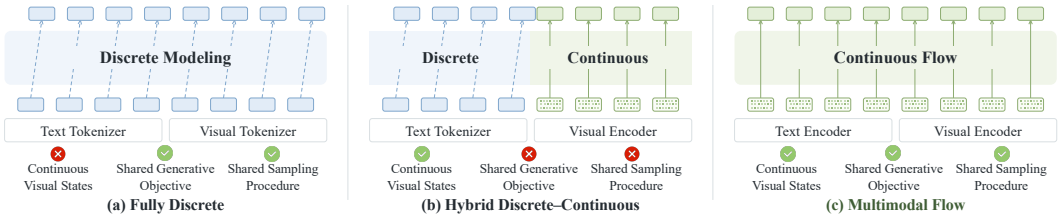

Figure 1: **Comparison of multimodal modeling paradigms.** Multimodal Flow combines continuous states with a shared generative objective and sampling procedure for language and vision. Blue and green denote discrete and continuous states respectively.

continuous modeling in embedding spaces, where language and vision can share a continuous objective and sampling mechanism while retaining modality-specific representations.

The technical ingredients for this paradigm are now available. Continuous diffusion and flow models are well established for visual generation (Esser et al., 2024; Lipman et al., 2022; Ho et al., 2020), and Embedded Language Flows (ELF) shows that contextualized text embeddings can also be modeled directly with Flow Matching (Hu et al., 2026). Prior multimodal diffusion and flow systems further demonstrate continuous generation along multiple image–text paths (Bao et al., 2023; Li et al., 2025a; He et al., 2025). These systems largely center on joint denoising or predefined cross-modal routes. A framework that causally factorizes task-defined multimodal sequences and learns from text-only, image-only, and bidirectional paired data through the same continuous pretraining objective remains underexplored.

A direct approach would concatenate continuous text and visual embeddings and apply a shared flow model. Concatenation alone, however, leaves three questions unanswered: what constitutes a generative unit for each modality, how those units form an ordered conditional process, and how the model can share cross-modal interaction while respecting modality-specific representation statistics. We address these questions with **Multimodal Flow**. First, modality-specific encoders map text blocks and images to continuous embeddings, forming hyperchunks that preserve text-token order and visual-grid structure, respectively. These heterogeneous hyperchunks provide common units of conditional generation that can be arranged into task-defined multimodal sequences. Second, a shared chunk-causal flow backbone conditions each target on preceding chunks while modeling positions within the target jointly. Third, joint attention supports cross-modal interaction, while modality-specific feed-forward networks adapt computation to each representation space. The corresponding decoders map generated states back to text or images. Multimodal Flow thus shares generative dynamics and cross-modal interaction without forcing language and vision into a homogeneous representation. Figure 3 shows parallel target-chunk prediction during training and sequential generation at inference. The same backbone and objective support mixed multimodal pretraining and finetuning for visual question answering and text-to-image generation.

We instantiate the framework as **MF-1** and evaluate its pretraining, competitive performance, and transfer. Across 0.6B–1.6B variants, longer pretraining and larger models generally improve language modeling, image captioning, and text-to-image generation. With 150B pretraining tokens, the 1.6B MF-1 achieves competitive generation and understanding performance against established unified models, scoring 0.821 on GenEval, 83.44 on DPG-Bench, and an average of 75.3 across VQAv2, MMBench and POPE. Controlled comparisons also show that Multimodal Flow outperforms fully discrete and hybrid architectures on the evaluated multimodal tasks. Separately, under matched total training-token budgets of the 1.6B architecture, mixed pretraining raises SeedBench from 31.6 to 62.4 and MMBench from 36.0 to 67.2 relative to a randomly initialized flow backbone, demonstrating MF-1's ability to transfer learned multimodal representations to downstream tasks. Additional analyses favor semantic visual representations and modality-specific feed-forward networks, and show that the same classifier-free guidance (CFG) (Ho & Salimans, 2022) mechanism can improve both text-to-image and image-conditioned text generation. Together, these results support the feasibility, competitiveness, and transferability of continuous chunk-based embedding modeling on the evaluated language–image tasks. Additional related work is discussed in Appendix A.

In summary, our main contributions are as follows:

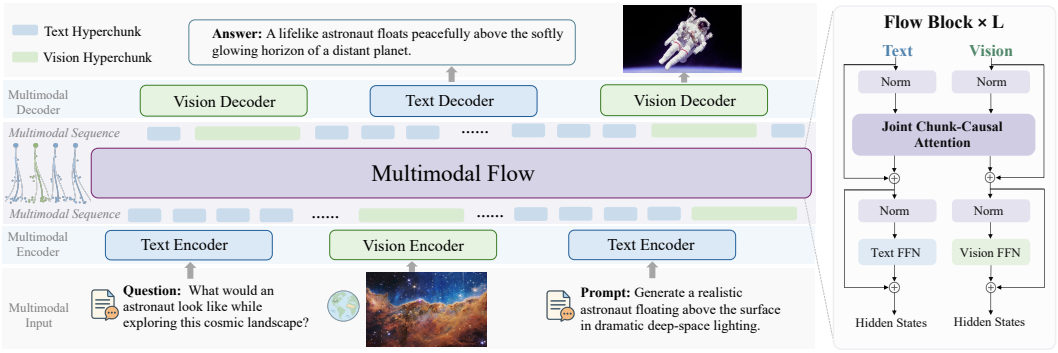

Figure 2: **Multimodal Flow architecture and ordered hyperchunk representation.** Frozen multimodal encoders map text blocks and images to continuous hyperchunks. After normalization and projection, a shared chunk-causal backbone applies joint attention and modality-specific FFNs. Multimodal decoders map generated hyperchunks back to text or images.

- We introduce Multimodal Flow, a fully continuous framework that models language and vision in their respective embedding spaces under one Flow Matching objective.
- We develop ordered hyperchunks and a chunk-causal backbone that preserve modality-specific structure, support task-defined multimodal sequences, and enable parallel target prediction followed by sequential generation.
- We validate MF-1 through scaling, matched architecture comparisons, and downstream transfer, demonstrating the strong multimodal modeling capabilities of Multimodal Flow.

2 Multimodal Flow

2.1 Overview

Figure 2 gives an overview of Multimodal Flow. The remainder of this section follows the model dataflow. We first describe the continuous representations and chunk construction, then introduce the chunk-causal architecture and Flow Matching objective, and finally present parallel training and sequential generation. Section 2.4 instantiates the same formulation for mixed multimodal pretraining and downstream finetuning.

2.2 Continuous Multimodal Representations

Multimodal Flow uses pretrained modality-specific representation encoders to map images and discrete text into continuous states. Given an image I and a text token sequence $Y = (y_1, \ldots, y_{L_\ell})$, we first partition the text into contiguous blocks of length B,

$$Y_b = Y_{(b-1)B+1:\min(bB, L_\ell)}, \tag{1}$$

and encode each block independently. The normalized visual and text representations are

$$\mathbf{x}^v = \mathrm{Norm}_v(E_v(I)), \qquad \mathbf{x}^\ell_b = \mathrm{Norm}_\ell(E_\ell(Y_b)). \tag{2}$$

Here, E_v and E_ℓ denote the visual and text encoders, while Norm_v and Norm_ℓ denote the corresponding normalization operations. The two representations retain their respective dimensionalities and internal structures. Modality-specific input projections subsequently map them to a common hidden space.

The continuous states are organized into hyperchunks according to the structure of each modality. Each image forms a visual chunk that preserves its spatial layout, and each independently encoded text block forms a text chunk:

$$\mathbf{c}^v = \mathbf{x}^v, \qquad \mathbf{c}^\ell_b = \mathbf{x}^\ell_b. \tag{3}$$

Text chunks retain the original token order. We use $B = 8$ in our main models.

After generation, the predicted representations are denormalized and decoded into images or text:

$$\widehat{I} = D_v\big(\mathrm{Norm}_v^{-1}(\widehat{\mathbf{x}}^v)\big), \qquad \widehat{Y}_b = D_\ell\big(\mathrm{Norm}_\ell^{-1}(\widehat{\mathbf{x}}_b^\ell)\big). \tag{4}$$

In our implementation, we use frozen SigLIP 2 (Tschannen et al., 2025) and T5-small (Raffel et al., 2020) encoders, together with a pretrained image decoder (Tong et al., 2026) and a separately trained text decoder. Both decoders remain frozen during flow pretraining. Representation and decoder details are provided in the appendix.

2.3 Chunk-Causal Multimodal Flow

2.3.1 Chunk-Causal Flow Modeling

Multimodal Flow represents language and vision as an ordered sequence of chunks,

$$\mathcal{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_K), \tag{5}$$

and factorizes their joint distribution according to the chunk order:

$$p_\theta(\mathcal{C}) = \prod_{k=1}^{K} p_\theta(\mathbf{c}_k \mid \mathbf{c}_{<k}). \tag{6}$$

When predicting chunk $\mathbf{c}_k$, all preceding chunks are visible and future chunks remain hidden, while positions within the current chunk are modeled jointly. The chunk order and target modality are specified by the task, allowing the same backbone to operate under different multimodal contexts.

Let $\mathbf{x}^{(k)}$ denote the clean representation of the target chunk. We construct a linear probability path (Lipman et al., 2022):

$$\mathbf{z}_t^{(k)} = t\mathbf{x}^{(k)} + (1-t)\boldsymbol{\epsilon}^{(k)}, \qquad \boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(0, \mathbf{I}), \tag{7}$$

where $t = 0$ corresponds to noise and $t = 1$ corresponds to the clean representation. Given the preceding chunks and the perturbed target state, the chunk-causal flow backbone predicts the clean endpoint:

$$\widehat{\mathbf{x}}_\theta^{(k)} = f_\theta^{m_k}\left(\mathbf{z}_t^{(k)}, t \mid \mathbf{c}_{<k}\right), \tag{8}$$

where m_k denotes the modality of the target chunk.

Visual and text states are mapped to a common hidden space through modality-specific input projections and augmented with modality embeddings and timestep embeddings. Multimodal rotary position embedding (MRoPE) (Wang et al., 2024a) encodes both the chunk order and the positional structure within each chunk. The resulting states are processed by joint self-attention under a chunk-causal mask, allowing information from all preceding visual and text chunks to contribute to the target prediction. The attention outputs are then processed by modality-specific feed-forward networks and projected back to their respective continuous representation spaces.

2.3.2 Parallel Flow Matching Training

During training, the chunk-causal mask allows multiple target chunks to be predicted in parallel. For a sequence containing multiple text blocks, we construct clean and perturbed views of each block. A perturbed target block can attend to all preceding clean chunks and its own perturbed state, but not to its clean counterpart or any future chunk. Multiple chunk predictions can therefore be computed in a single forward pass. Each target chunk receives an independently sampled flow timestep. Figure 3(a) illustrates this parallel construction as part of the overall training procedure.

The model predicts the clean endpoint and converts it into the corresponding velocity along the probability path:

$$\widehat{\mathbf{v}}_\theta^{(k)} = \frac{\widehat{\mathbf{x}}_\theta^{(k)} - \mathbf{z}_t^{(k)}}{1-t}, \qquad \mathbf{v}^{(k)} = \mathbf{x}^{(k)} - \boldsymbol{\epsilon}^{(k)}. \tag{9}$$

Let $\mathcal{K}$ denote the set of target chunks predicted in parallel, and let $\mathbf{M}^{(k)}$ denote the valid-position mask for chunk k. The training objective is

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\frac{1}{\sum_{k\in\mathcal{K}}|\mathbf{M}^{(k)}|}\sum_{k\in\mathcal{K}}\sum_i M_i^{(k)}\left\|\widehat{\mathbf{v}}_{\theta,i}^{(k)} - \mathbf{v}_i^{(k)}\right\|_2^2\right]. \tag{10}$$

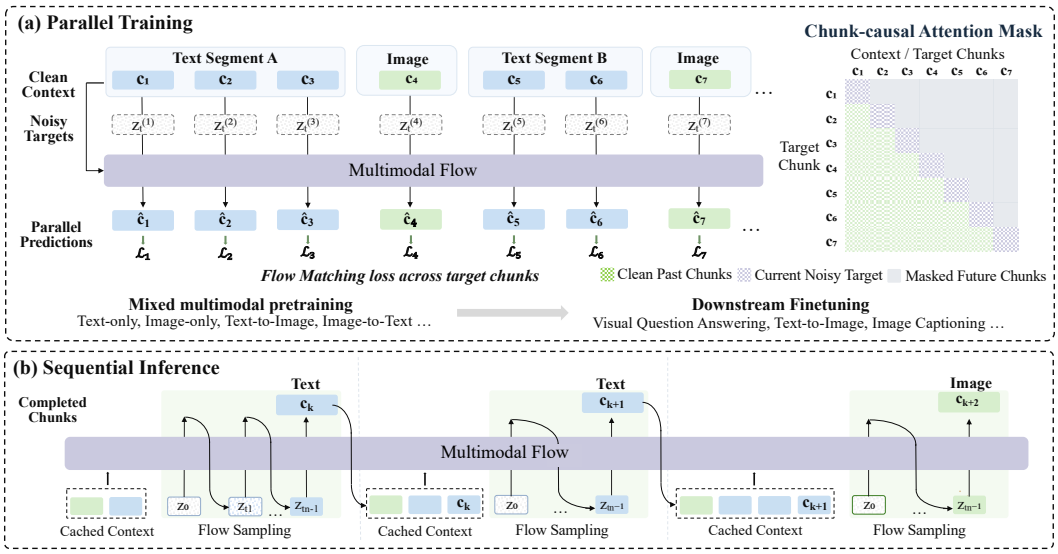

Figure 3: **Parallel training, mixed multimodal pretraining, and sequential inference.** The chunk-causal formulation predicts multiple target chunks in parallel during training and generates chunks sequentially at inference. Different chunk sequences express unimodal modeling, cross-modal generation, and downstream finetuning within the same Flow Matching objective.

To improve computational efficiency for variable-length multimodal sequences, we use sequence packing to place multiple logical chunk sequences into a packed sequence with a fixed maximum length. The chunk-causal mask uses sequence identifiers to isolate different samples and prevent information exchange across sequence boundaries. Details of timestep sampling, endpoint handling, and sequence packing are provided in the appendix.

2.3.3 Sequential Chunk Inference

As illustrated in Figure 3(b), the model generates chunks sequentially in the order specified by the task. Given a clean chunk prefix, the next target chunk is initialized from Gaussian noise. The model then integrates the learned vector field from $t = 0$ to $t = 1$, conditioned on the preceding chunks. Once generated, the chunk is appended to the context before the model proceeds to the next chunk.

Chunk-causal attention ensures that the hidden states of completed chunks do not depend on future chunks, which enables KV caching. The keys and values of preceding clean chunks are cached at each layer. At each sampling step, only the current target chunk needs to be recomputed. Once generated, its clean state is added to the cache for subsequent chunks.

For conditional generation, we apply CFG. Let $\widehat{\mathbf{x}}_{\theta,c}$ and $\widehat{\mathbf{x}}_{\theta,\emptyset}$ denote the predictions obtained with and without the conditioning chunks, respectively. The guided prediction is

$$\widehat{\mathbf{x}}_{\theta,\gamma} = \widehat{\mathbf{x}}_{\theta,\emptyset} + \gamma \left(\widehat{\mathbf{x}}_{\theta,c} - \widehat{\mathbf{x}}_{\theta,\emptyset} \right), \tag{11}$$

where γ is the guidance scale. By specifying different chunk prefixes and target modalities, the same sequential generation process supports both unimodal and cross-modal inference.

2.4 Mixed Multimodal Pretraining and Downstream Finetuning

2.4.1 Mixed Multimodal Pretraining

The chunk representation converts heterogeneous data sources into task-defined sequences with a common target interface. For text-only data, consecutive text blocks form a sequence $(\mathbf{c}_1^{\ell}, \ldots, \mathbf{c}_K^{\ell})$, and every chunk is a target conditioned on its clean prefix; the first chunk is therefore unconditional. An image-only example contains a target visual chunk with no conditioning chunk. For paired image–text data, we use both modality orders: text chunks followed by a visual target for text-to-image generation, and a visual chunk followed by text targets for image-to-text generation. Figure 3 lists these training configurations.

We pretrain one chunk-causal flow backbone on a mixture of these tasks. Let τ be a task sampled from mixture distribution π, let $s \sim \mathcal{D}_\tau$ be a sample from its data source, and let $\mathcal{G}_\tau(s)$ construct the corresponding chunk sequence and target set. The mixed-pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathbb{E}_{\tau\sim\pi,\, s\sim\mathcal{D}_\tau}\left[\mathcal{L}_{\text{flow}}(\mathcal{G}_\tau(s))\right]. \tag{12}$$

Across tasks, the continuous interfaces and flow backbone remain fixed under the same Flow Matching objective. Only chunk content, ordering, and target modalities vary. The resulting mixture learns unimodal distributions and bidirectional cross-modal conditionals within a single model.

2.4.2 Downstream Finetuning

Downstream finetuning changes the task sequence and data, but not the model interface or objective. For visual question answering, a visual chunk and question chunks form the clean prefix, while answer chunks are text targets. For text-to-image generation, prompt chunks form the prefix and the image is the visual target. The frozen modality codecs are reused, and the chunk-causal backbone is initialized from mixed pretraining and optimized with the same $\mathcal{L}_{\text{flow}}$.

This separation between task specification and generative modeling is central to Multimodal Flow. Task-specific data and optimization adapt the pretrained backbone, while the chunk interface, causal factorization, and continuous objective remain the same across understanding and generation tasks.

3 Experiments

3.1 Experimental Setup

We evaluate MF-1 on language modeling, image captioning, multimodal understanding, and image generation. Our main model has a 1.6B parameter flow backbone trained from scratch with 150B tokens. All MF-1 results in Tables 1–3 use the same checkpoint after 5B tokens of joint finetuning. We also compare 0.6B, 1.2B, and 1.6B variants under a common pretraining and evaluation protocol. Appendices B and C provide the model and codec configurations, training budgets, benchmarks, and evaluation sampling settings.

3.2 Comparison across Multimodal Modeling Paradigms

We evaluate MF-1 through two complementary comparisons. The first uses published results from generation-specific, understanding-specific, and unified models under their reported training settings. The second compares modeling paradigms with matched data and optimization schedules under the same trainable-parameter budget.

The image-generation comparisons show that MF-1 achieves the strongest compositional generation performance and the best long-prompt alignment among unified models. This indicates that jointly modeling language and images within a continuous generative process produces language representations that transfer effectively to text-conditioned image synthesis. MF-1 also performs strongly on multimodal understanding despite being trained from scratch on only 150B tokens. It consistently surpasses comparable from-scratch unified models such as Muddit and D-DiT, while remaining competitive with similarly sized models initialized from pretrained LLMs.

To assess Multimodal Flow under a matched training budget, we compare it with fully discrete and hybrid architectures using the same data, optimization schedule, and trainable-parameter budget. The Transfusion-style hybrid provides a particularly close reference, sharing MF-1's visual pathway and backbone design while using autoregressive text modeling. MF-1 achieves stronger image generation and visual understanding across these comparisons (Table 4), establishing it as an effective fully continuous architecture for unified multimodal modeling.

3.3 Mixed Multimodal Pretraining and Model Capacity

Figure 4 tracks the 0.6B, 1.2B, and 1.6B variants under the same mixed-pretraining and evaluation protocol, reporting the Flow Matching objective, language PPL, GenEval, and image-captioning CLIPScore. All three variants show consistent overall improvement: the Flow Matching objective

Table 1: Category-level comparison on GenEval. Unified models are grouped by multimodal modeling paradigm. Best and second-best results are shown in bold and underlined respectively.

Type	Model	Params	Single	Two	Count.	Colors	Pos.	Color Attr.	Overall↑
Gen. only	SDXL (Podell et al., 2024)	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	Hunyuan-DiT (Li et al., 2024b)	1.5B	0.97	0.77	0.71	0.88	0.13	0.30	0.63
	SD3 Medium (Esser et al., 2024)	2.0B	0.98	0.74	0.63	0.67	0.34	0.36	0.62
Unified	*Fully discrete*								
	LWM (Liu et al., 2025a)	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	Show-o (Xie et al., 2025)	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Emu3-Gen (Wang et al., 2024b)	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	Muddit (Shi et al., 2025)	1B	0.98	0.72	0.54	0.82	0.19	0.41	0.61
	UniDisc (Swerdlow et al., 2025)	1.4B	0.92	0.47	0.15	0.67	0.13	0.19	0.42
	Hybrid discrete–continuous								
	Transfusion (Zhou et al., 2025)	7B	–	–	–	–	–	–	0.67
	JanusFlow (Ma et al., 2025)	1.3B	0.97	0.59	0.45	0.83	0.53	0.42	0.63
	D-DiT (Li et al., 2025b)	2B	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	Fully continuous								
	MF-1	1.6B	**0.99**	**0.92**	**0.73**	**0.90**	**0.81**	**0.58**	**0.82**

Table 2: Detailed text-to-image generation performance on DPG-Bench.

Model	Params	Global	Entity	Attribute	Relation	Other	Overall↑
SDXL (Podell et al., 2024)	2.6B	83.27	82.43	80.91	86.76	80.41	74.65
Hunyuan-DiT (Li et al., 2024b)	1.5B	84.59	80.59	88.01	74.36	86.41	78.87
SD3 Medium (Esser et al., 2024)	2.0B	87.90	**91.01**	88.83	80.70	88.68	**84.08**
Show-o (Xie et al., 2025)	1.3B	79.33	75.44	78.02	84.45	60.80	67.27
Janus (Wu et al., 2025)	1.3B	82.33	87.38	87.70	85.46	86.41	79.68
Emu3-Gen (Wang et al., 2024b)	8B	85.21	86.68	86.84	**90.22**	83.15	80.60
JanusFlow (Ma et al., 2025)	1.3B	87.03	87.31	87.39	89.79	88.10	80.09
MF-1	1.6B	**89.38**	89.03	**90.82**	89.73	**89.47**	83.44

Table 3: Multimodal understanding performance and pretraining scale. PT Tok. denotes cumulative pretraining tokens along the backbone inheritance chain.

Paradigm	Model	Params	PT Tok.	POPE	MMB	SEEDB	VQAv2	GQA	OK-VQA
With Pretrained LLM Initialization									
Und. only	MobileVLM-V2 (Chu et al., 2024)	2.7B	∼1.3T	84.7	63.2	–	–	61.1	–
	LLaVA-Phi (Zhu et al., 2024)	2.7B	∼1.4T	85.0	59.8	–	71.4	–	–
	LLaVA-v1.5 (Liu et al., 2024a)	7.0B	∼2.0T	85.9	64.3	58.6	78.5	62.0	–
	Qwen-VL-Chat (Bai et al., 2023)	7.0B	∼3.1T	83.7	60.6	58.2	78.2	57.5	56.6
Discrete	Janus (Wu et al., 2025)	1.3B	∼0.9T	87.0	69.4	63.7	77.3	59.1	–
	Show-o (Xie et al., 2025)	1.3B	∼1.0T	73.8	–	–	59.3	48.7	–
	Janus-Pro (Chen et al., 2025c)	1.5B	∼1.1T	86.2	75.5	68.3	–	59.3	–
Hybrid	JanusFlow (Ma et al., 2025)	1.3B	∼2.0T	88.0	74.9	70.5	79.8	60.3	–
	Show-o2 (Xie et al., 2026)	1.5B	∼18.2T	83.0	67.4	65.6	–	60.0	–
Without Pretrained LLM Initialization									
Discrete	Muddit (Shi et al., 2025)	1.0B	∼400B	52.5	28.4	13.0	68.2	57.5	21.1
Hybrid	D-DiT (Li et al., 2025b)	2.0B	∼4.3T	79.2	–	–	59.5	55.1	28.5
Continuous	**MF-1**	1.6B	∼150B	86.1	67.2	62.4	72.6	58.3	39.1

and language PPL decrease, while GenEval and CLIPScore increase as training proceeds. Increasing model capacity yields clearer gains in language modeling and image captioning, with the 1.6B model maintaining the strongest performance later in training. These trends show that a shared chunk-causal Flow Matching objective can jointly develop language modeling, image-conditioned text

Table 4: Controlled comparison of multimodal architectures under matched data, optimization schedules, and trainable parameter budgets.

Architecture	Architecture Type	GenEval ↑	GQA ↑	VQAv2 ↑	MMBench ↑	SEEDB ↑
Multimodal Flow	Fully Continuous	**0.7134**	**55.60**	**69.03**	**46.74**	**51.85**
Transfusion-Style	Discrete–Continuous Hybrid	0.6693	52.81	68.49	33.68	31.31
Chameleon-Style	Fully discrete	0.3744	45.83	56.37	38.40	42.40

Table 5: Downstream performance with random and mixed-pretrained initialization.

Initialization	GenEval ↑	DPG-Bench ↑	SEEDB ↑	MMB ↑	OK-VQA ↑
Random	0.527	57.81	31.6	36.0	22.9
Mixed pretraining	**0.821**	**83.44**	**62.4**	**67.2**	**39.1**

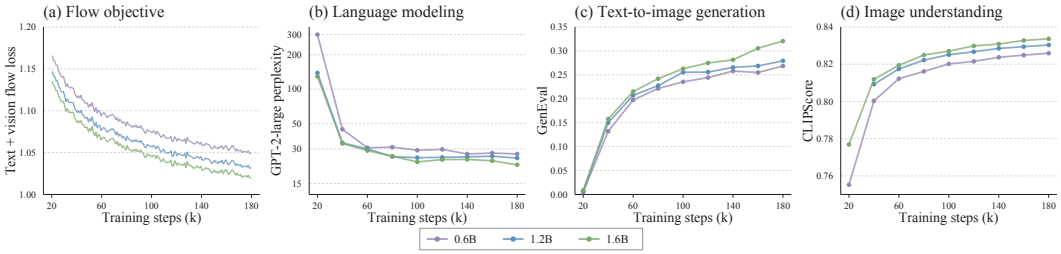

Figure 4: Training progress under mixed multimodal pretraining at different model capacities. We compare the Flow Matching objective, GPT-2-large PPL, GenEval, and CLIPScore for the 0.6B, 1.2B, and 1.6B models.

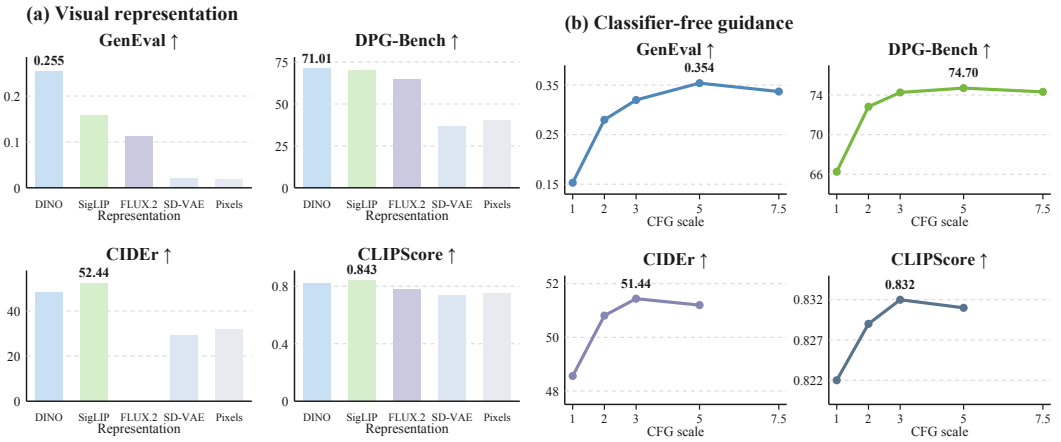

Figure 5: Design analysis. (a) Image-generation and image-captioning performance under different visual representation spaces. (b) Performance as a function of classifier-free guidance scale. Outlined marks and value labels indicate the best result for each metric.

generation, and text-to-image generation capabilities, with further gains from longer training and increased model capacity.

3.4 Downstream Finetuning from Mixed Pretraining

We compare mixed-pretrained and randomly initialized 1.6B models with the same architecture, matching the baseline's downstream training tokens to the pretrained model's combined pretraining and finetuning budget. Table 5 shows consistent gains in GenEval, DPG-Bench and all VQA benchmarks, supporting the transfer benefits of mixed pretraining rather than increased token exposure.

Table 6: Comparison of cross-modal parameterizations after 50B pretraining tokens.

Attn. projection	FFN	Text PPL ↓	GenEval ↑	DPG ↑	CIDEr ↑	CLIPScore ↑
Specific	Specific	28.09	**0.255**	71.01	**48.14**	**0.820**
Specific	Shared	29.67	0.233	70.82	47.31	**0.820**
Shared	Specific	**27.43**	0.226	**71.04**	47.91	**0.820**
Shared	Shared	28.35	0.237	69.70	42.02	0.788

3.5 Design Analysis

We analyze three design choices in MF-1: the continuous visual representation space, the balance between shared and modality-specific computation, and CFG for text and image generation.

3.5.1 Continuous Visual Representation Space

The visual representation encoder defines the continuous state modeled by the flow. Figure 5(a) compares DINOv2 (Oquab et al., 2023) and SigLIP2 (Tschannen et al., 2025) representations with FLUX.2 VAE latents (Black Forest Labs, 2025), SD-VAE latents (Rombach et al., 2022), and raw image patches.

At the evaluated training budget, DINOv2 achieves the highest image-generation scores, whereas SigLIP2 provides a better balance between generation and captioning, yielding higher CIDEr and CLIPScore despite lower GenEval and DPG-Bench scores. Reconstruction-oriented VAE representations underperform semantic embeddings on both generation and captioning metrics in this setting. Raw pixel representations also yield weak performance on both tasks.

3.5.2 Shared Interaction and Modality-Specific Computation

MF-1 uses joint attention for cross-modal interaction, with either shared or modality-specific attention projections and FFNs. Table 6 compares four parameter-matched configurations. Shared and modality-specific attention projections perform similarly overall, whereas sharing FFNs degrades performance, particularly on image-conditioned text generation. These results motivate shared attention projections for cross-modal interaction and modality-specific FFNs to accommodate the distinct representation statistics of language and vision.

3.5.3 Bidirectional Classifier-Free Guidance

MF-1 applies the same CFG mechanism to text-to-image and image-conditioned text generation. We fix the pretrained EMA checkpoint of the 0.6B model and vary only the guidance scale during inference.

As shown in Figure 5(b), image captioning performs best at a guidance scale of 3, while text-to-image generation reaches its strongest performance at a scale of 5 and begins to decline as the scale increases further. These results show that the same continuous flow formulation supports CFG for both text and image generation, with the guidance strength adjusted according to the output modality.

4 Conclusion

We introduced **Multimodal Flow**, a fully continuous model for language–vision pretraining, instantiated as **MF-1**. It organizes text blocks and images as ordered hyperchunks and models them with a shared chunk-causal flow backbone, preserving modality-specific structure within a common generative process. The resulting framework expresses diverse language–vision tasks as ordered hyperchunk sequences, unifying multimodal understanding and generation under a single continuous objective. Experiments show gains with increased model scale and training tokens, competitive generation and understanding performance, and substantial downstream benefits from mixed pretraining. Together, these results establish continuous hyperchunk modeling as a flexible paradigm for multimodal pretraining and show its ability to learn transferable representations across language and vision. Extending it to longer interleaved sequences, video, and other structured modalities remains a promising direction. We leave it as future work.

AI USE STATEMENT

We use generative AI tools to suggest and refine wording in selected passages, improve English expression, identify potentially relevant references, and provide feedback on figure design and the presentation of experimental comparisons. We have reviewed and edit all AI-assisted material, independently verified the cited literature and technical claims, and take full responsibility for the final manuscript and artifacts.

REPRODUCIBILITY STATEMENT

The main text and Appendices B and C describe the MF-1 architecture, representation and decoder interfaces, training objective, data sources and task mixture, training-token budgets, benchmarks, and inference settings. The controlled architecture comparison matches data exposure, optimization settings, and trainable-parameter budgets while retaining each architecture's native representations and objectives. We will release the code and model weights to facilitate independent reproduction.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

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.

Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison. https://bfl.ai/research/representation-comparison, November 2025. Technical report.

Keshigeyan Chandrasegaran, Kyle Sargent, Suchir Agarwal, Michael Jang, Michael Poli, Juan Carlos Niebles, Justin Johnson, Jiajun Wu, and Li Fei-Fei. Gpic: A giant permissive image corpus for visual generation. *arXiv preprint arXiv:2605.30341*, 2026.

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.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

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*, 2025c.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

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.

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.

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

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.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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.

Jiahao Guo, Sinan Du, Jingfeng Yao, Wenyu Liu, Bo Li, Haoxiang Cao, Kun Gai, Chun Yuan, Kai Wu, and Xinggang Wang. Visual generation tuning. *arXiv preprint arXiv:2511.23469*, 2025.

Ju He, Qihang Yu, Qihao Liu, and Liang-Chieh Chen. Flowtok: Flowing seamlessly across text and image tokens. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–12. IEEE, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024a.

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.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024b.

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.

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.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *International Conference on Learning Representations*, volume 2025, pp. 45953–45977, 2025a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2755–2765. IEEE, 2025b.

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, 2024b.

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.

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.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3190–3199. IEEE, 2019.

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

OpenDatasets. LAION DALL-E 3 Discord Dataset. Hugging Face dataset, https://huggingface.co/datasets/OpenDatasets/dalle-3-dataset.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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.

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.

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

Bao Tang, Jiahao Guo, Haoxiang Cao, Wenyu Liu, Changqian Yu, Kun Gai, and Xinggang Wang. Stablevq: Practical guidelines for stable vector-quantized tokenizer training, 2026.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

Hongyuan Tao, Bencheng Liao, Shaoyu Chen, Haoran Yin, Qian Zhang, Wenyu Liu, and Xinggang Wang. Infinitevl: Synergizing linear and sparse attention for highly-efficient, unlimited-input vision-language models. *arXiv preprint arXiv:2512.08829*, 2025.

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

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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Ultra-FineWeb: Efficient data filtering and verification for high-quality llm training data, 2025.

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, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *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.

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.

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, 2026.

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, and Xinggang Wang. Diffusionvl: Translating any autoregressive models into diffusion vision language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2026.

Shuai Zhang, Bao Tang, Siyuan Yu, Yueting Zhu, Jingfeng Yao, Ya Zou, Shanglin Yuan, Li Yu, Wenyu Liu, and Xinggang Wang. Mobilei2v: Fast and high-resolution image-to-video on mobile devices. *arXiv preprint arXiv:2511.21475*, 2025.

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.

Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, and Xinggang Wang. Dig: Scalable and efficient diffusion models with gated linear attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7664–7674. IEEE, 2025.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st international workshop on efficient multimedia computing under limited*, pp. 18–22, 2024.

Yueting Zhu, Yuehao Song, Kaicheng Zhang, Bao Tang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Stream forcing: Constructing unified training trajectory for robust streaming video generation. *arXiv preprint arXiv:2608.10439*, 2026.

Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv preprint arXiv:2503.08686*, 2025.

A Related Work

A.1 Discrete and Hybrid Discrete–Continuous Multimodal Modeling

Generative foundation models for language and vision have largely followed two architectural paradigms: fully discrete and hybrid discrete–continuous modeling. Fully discrete models use a visual tokenizer to map images into discrete visual tokens. Text and images can then form a common discrete sequence and be modeled through autoregressive prediction, masked diffusion, or discrete flow (Team, 2024; Wang et al., 2024b; Xie et al., 2025; Yang et al., 2026; Wang et al., 2026). This design aligns the state type, prediction target, and generative process across modalities, while allowing the model to reuse established language model training recipes. For images, however, representation quality is bounded by the visual tokenizer. Improving reconstruction fidelity generally requires a larger codebook or longer token sequences, while information lost through quantization cannot be fully recovered by the downstream backbone.

Hybrid discrete–continuous models preserve discrete language modeling while generating continuous visual representations with diffusion or flow. A representative design retains autoregressive next-token prediction for text and pairs it with a continuous generation objective for images (Zhou et al., 2025; Deng et al., 2025; Ma et al., 2025; Guo et al., 2025). Other systems replace the autoregressive language objective with discrete diffusion or related flow-based objectives (Li et al., 2025b; Nguyen et al., 2025). These models avoid visual quantization, but a single system must still coordinate different training objectives, perturbation schedules, and sampling procedures.

A.2 Continuous Generative Modeling of Language and Vision

Continuous diffusion and flow models are now widely used for visual generation. They typically learn probability paths from noise to data in reconstruction-oriented VAE latent spaces (Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2022; Zhu et al., 2025; Zhang et al., 2025; Zhu et al., 2026). Representation Autoencoders further incorporate pretrained visual encoders, producing continuous generative states with richer semantic information and spatial structure (Zheng et al., 2026; Tong et al., 2026). More recently, continuous generative modeling has been extended to language. One line of work operates directly in token-level embedding spaces, while another compresses sentences or text blocks into continuous latents and recovers discrete text with a language decoder (Li et al., 2022; Gong et al., 2022; Lovelace et al., 2023; Hu et al., 2026; Guo et al., 2026). Together, these studies demonstrate continuous language generation across different representation granularities, ranging from token-level embeddings to compressed sentence- and block-level latents.

Prior studies have established continuous multimodal generation through composable diffusion, joint denoising, and cross-modal flows (Xu et al., 2023; Tang et al., 2023; Bao et al., 2023; Li et al., 2025a; Liu et al., 2025b; He et al., 2025). Many nevertheless generate from reconstruction-oriented VAE latents or compact visual tokens rather than semantic representations. These approaches also tend to couple modalities through joint denoising or direct mappings, without explicitly treating token-ordered text and spatially structured images as units of an ordered conditional sequence. Meanwhile, mixed pretraining of a randomly initialized continuous backbone on multimodal data remains largely unexplored. Multimodal Flow addresses these gaps by organizing contextual text embeddings and spatial semantic visual embeddings into ordered hyperchunks and modeling them with one chunk-causal Flow Matching objective. MF-1's scaling, transfer, and controlled comparisons show that this formulation can serve as a practical foundation for multimodal pretraining beyond cross-modal generation.

B Implementation Details

B.1 Backbone Configuration

Table 7 summarizes the main 1.6B MF-1 configuration.

Table 7: Backbone and representation settings of the main MF-1 model.

Backbone	Setting	Representation	Setting
Parameters	1.6B	Text encoder	T5-small
Transformer layers	32	Text embedding dim.	512
Hidden dimension	1600	Tokens per text chunk	8
Attention heads	25	Text input bottleneck	128
Head dimension	64	Vision encoder	SigLIP2-so400m
FFN hidden dimension	4224	Image input resolution	224
Attention projections	Shared	Visual grid	16×16
FFN parameters	Modality-Specific	Visual embedding dim.	1152

B.2 Representation Encoders and Decoders

Text. We pretrain the text decoder separately from the multimodal flow backbone. A frozen T5-small encoder maps text to 512-dimensional latents, which are normalized and detached before decoder training. For each token position, we independently sample $z \sim \mathcal{N}(0,1)$, set $\lambda = \sigma(0.8 + 0.8z)$, and perturb the clean latent h as $\widetilde{h} = \lambda h + (1-\lambda)s\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and s controls the noise strength. A six-layer bidirectional Transformer of width 512 predicts the original tokens in parallel within each eight-token block. We minimize cross-entropy averaged over target text tokens, excluding conditioning and padding positions. This stage trains only the decoder. The resulting decoder is frozen during multimodal pretraining. At inference, it converts generated text latents into tokens block by block.

Vision. The main configuration uses a frozen SigLIP2-so400m encoder (Tschannen et al., 2025) with a patch size of 14. Each 224×224 image produces 256 patch embeddings of dimension 1152, retained as a single spatially structured chunk. A pretrained representation decoder (Tong et al., 2026) reconstructs images from the generated visual embeddings. Both text and visual codecs remain frozen during flow pretraining and downstream finetuning.

Normalization. We normalize both modalities before flow modeling. For each active text token, the frozen T5-small encoder produces $h^{\text{text}}_{i,d}$, which is normalized as $\tilde{h}^{\text{text}}_{i,d} = (h^{\text{text}}_{i,d} - \mu_{\text{text}})/\sigma_{\text{text}}$ using dimension-shared constants $\mu_{\text{text}} = 0$ and $\sigma_{\text{text}} = 0.2$. EOS tokens are treated identically; padding positions are set to zero and excluded from the objective. For vision, we normalize each token position and channel: $\tilde{h}^{\text{vis}}_{i,d} = (h^{\text{vis}}_{i,d} - \mu^{\text{vis}}_{i,d})/\sigma^{\text{vis}}_{i,d}$. The 256×1152 population moments were estimated from 50,176 GPIC training images. Generated vision states are denormalized before image decoding, $h^{\text{vis}}_{i,d} = \tilde{h}^{\text{vis}}_{i,d}\sigma^{\text{vis}}_{i,d} + \mu^{\text{vis}}_{i,d}$.

B.3 Flow Training and Sequence Packing

Timestep sampling and endpoint conversion. We sample a noise level from a shifted logit-normal distribution: $z \sim \mathcal{N}(0,1)$, $u = \text{sigmoid}(z)$, $q = \alpha u/[1 + (\alpha - 1)u]$, and $t = 1 - q$. We set $\alpha = 8$ for image targets and $\alpha = 6$ for text targets. Timesteps are sampled independently for target chunks: one per 256-token image chunk and one per 8-token text block. With x denoting the clean normalized state, the flow path is $z_t = tx + (1-t)\epsilon$. For numerical stability near the clean endpoint, the implementation converts the predicted clean state to velocity using $\hat{v}_\theta = (\hat{x}_\theta - z_t)/\max(1-t, 0.05)$.

Sequence packing. We pack independent examples into physical sequences of at most 32,768 model positions. Each example retains a distinct sequence identifier: chunk-causal attention permits

the prescribed interactions within an example but blocks attention across examples in the same pack. Padding positions are inactive in attention.

C Experimental Protocols

C.1 Training Configurations

We distinguish the main benchmark model from the controlled architecture and design studies. Table 8 records their model sizes and pretraining and finetuning token budgets. The scaling study follows the 0.6B, 1.2B, and 1.6B variants at matched checkpoints, as shown in Figure 4.

Table 8: Training token budgets for the main and controlled experiments. PT and FT denote pretraining and finetuning respectively.

Experiment	Reference	Model size	PT Tokens	FT Tokens
Main benchmark comparisons	Tables 1–3	1.6B	150B	5B
Controlled architectures	Table 4	1.6B	50B	5B
Pretraining transfer	Table 5	1.6B	150B	5B
Attention/FFN parameterization	Table 6	0.6B	50B	0B
Visual representations and CFG	Figure 5	0.6B	50B	0B

Pretraining data. We use GPIC (Chandrasegaran et al., 2026) for image generation, LLaVA-OneVision-1.5 (An et al., 2025) for image understanding, and Ultra-FineWeb-L3 (Wang et al., 2025) for text-only modeling. The image-only task uses GPIC images without their paired text.

Finetuning data. For text-to-image finetuning, we use BLIP3o-60k (Chen et al., 2025a), the LAION DALL-E 3 Discord dataset (OpenDatasets), and ShareGPT-4o-Image (Chen et al., 2025b). For VQA finetuning, we primarily use the LLaVA-v1.5 visual instruction-tuning data (Liu et al., 2024a).

Mixed pretraining. The task mixture consists of 70% text-only modeling, 20% image understanding, 9% text-to-image generation, and 1% image-only modeling. The main 1.6B model is pretrained from scratch on 150B tokens, followed by 5B finetuning tokens for downstream evaluation. The same Flow Matching objective is used across unimodal and cross-modal tasks.

Controlled comparisons. The architecture comparison uses 1.6B models with 50B pretraining and 5B finetuning tokens under matched data and optimization settings. The pretraining-transfer comparison instead uses a 1.6B model with 150B pretraining and 5B finetuning tokens. Its randomly initialized counterpart is trained directly on downstream data for 155B tokens, matching the total token budget. The attention/FFN and visual representation/CFG studies use 0.6B models with 50B pretraining without downstream finetuning.

Architecture-specific configurations. Table 9 summarizes the representations, training objectives, and parameter-sharing choices of the architectures compared in Table 4.

All trainable backbones are randomly initialized and share 32 layers, width 1600, 25 attention heads of dimension 64, RMSNorm, and positional encoding conventions. Their FFN widths are adjusted so that trainable parameter counts differ from MF-1's 1,634,591,168 by at most 25,536 (0.00156%); frozen codecs are excluded. The comparison matches the data stream, supervised-token exposure, optimization recipe, and trainable-parameter budget, while preserving each architecture's native objective.

C.2 Benchmarks and Metrics

Language modeling and captioning. We evaluate generated text using GPT-2-large perplexity (Radford et al., 2019), an external fluency metric rather than the flow model's likelihood. Image

Table 9: Representations and objectives in the controlled architecture comparison. AR denotes autoregressive cross-entropy; flow denotes velocity MSE. Parenthesized values in the last column are FFN hidden dimensions.

Architecture	Text	Vision	Attention / FFNs
Multimodal Flow	T5 latent (512; 8-token blocks), flow	SigLIP2 latent (256×1152), flow	Shared / modality-specific (4224)
Chameleon-Style	T5 token IDs, AR	Janus VQ-16 (576 IDs; 16,384 codes), AR	Shared / modality-specific (4000)
Transfusion-style	T5 token IDs, AR	SigLIP2 latent (256×1152), flow	Shared / modality-specific (4062)

captioning is evaluated on the COCO Karpathy test split (Lin et al., 2014; Karpathy & Fei-Fei, 2015) using CIDEr (Vedantam et al., 2015) and CLIPScore (Hessel et al., 2021).

Multimodal understanding. We report POPE (Li et al., 2023), MMBench (Liu et al., 2024b), SEED-Bench (Li et al., 2024a), VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), and OK-VQA (Marino et al., 2019).

Image generation. GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024) assess compositional generation and long-prompt adherence, respectively. Lower values are better for perplexity; higher values are better for the other reported metrics.

C.3 Evaluation Sampling

Table 10 gives the evaluation settings for visual question answering and text-to-image generation. For text outputs, the sampling steps apply to each generated chunk. The CFG sweep in Figure 5(b) varies the guidance scale instead of fixing it to the values below.

Table 10: Inference settings for visual question answering and image generation.

Task	Sampler	Steps	CFG scale	SDE γ
Visual question answering	SDE	16	3.0	1.0
Text-to-image generation	ODE	64	5.0	–