Think in Strokes, Not Pixels: Process-Driven Image Generation via Interleaved Reasoning

Lei Zhang[1,2] ⋆, Junjiao Tian[2], Zhipeng Fan[2], Kunpeng Li[2], Jialiang Wang[2], Weifeng Chen[2], Markos Georgopoulos[2], Felix Juefei-Xu[2], Manling Li[3] †, Julian McAuley[1] †, and Zecheng He[2] †

[1] University of California San Diego
[2] Meta Superintelligence Lab
[3] Northwestern University

Abstract. Humans paint images incrementally: they plan a global layout, sketch a coarse draft, inspect, and refine details, and most importantly, each step is grounded in the evolving visual states. However, can unified multimodal models trained on text-image interleaved datasets also imagine the chain of intermediate states? In this paper, we introduce **process-driven image generation**, a multi-step paradigm that decomposes synthesis into an **interleaved reasoning** trajectory of thoughts and actions. Rather than generating images in a single step, our approach unfolds across multiple iterations, each consisting of 4 stages: textual planning, visual drafting, textual reflection, and visual refinement. The textual reasoning explicitly conditions how the visual state should evolve, while the generated visual intermediate in turn constrains and grounds the next round of textual reasoning. A core challenge of process-driven generation stems from the ambiguity of intermediate states: how can models evaluate each partially-complete image? We address this through dense, step-wise supervision that maintains two complementary constraints: for the visual intermediate states, we enforce the spatial and semantic consistency; for the textual intermediate states, we preserve the prior visual knowledge while enabling the model to identify and correct prompt-violating elements. This makes the generation process explicit, interpretable, and directly supervisable. To validate proposed method, we conduct experiments under various text-to-image benchmarks.

Keywords: Interleaved Reasoning · Unified Multimodal Model

1 Introduction

Despite impressive progress in image generation, today's models still remain brittle on elementary visual logic and produce plausible but incorrect images. As shown in Fig. 1, a prompt of a bear *hovering above* a spoon might incorrectly yield a bear *standing beside* it. Such a one-shot black-box generation

† Joint last author.
⋆ Work done at Meta.

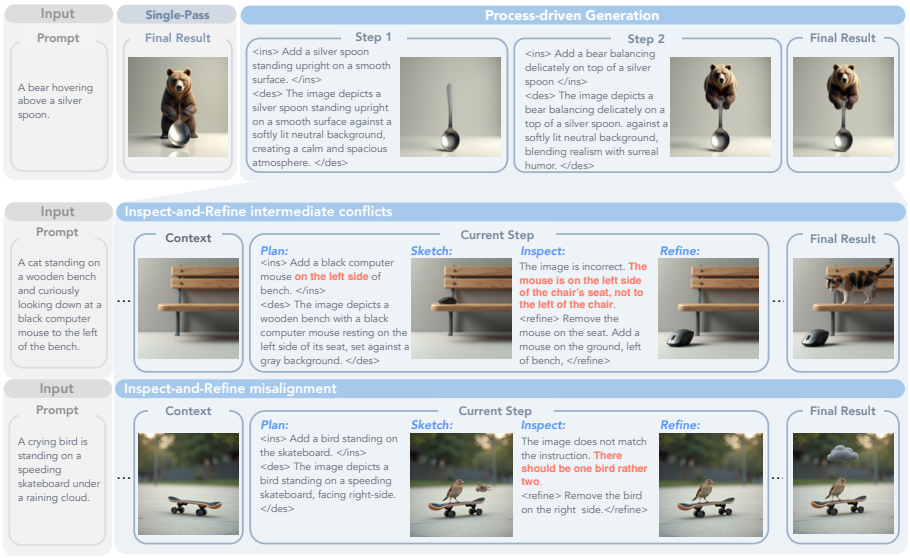

Fig. 1: Single-pass generation vs process-driven interleaved reasoning. Instead of training model to generate a final image within a single pass, we teach a unified multimodal foundation model to construct the image stroke by stroke, decision by decision. This process of `Plan`, `Sketch`, `Inspect`, and `Refine` enables transformation of ambiguous intermediates into compositionally faithful final images.

forces the model to commit to an entire scene within a single forward pass, resolving precise spatial layouts, object relations, and fine-grained attributes. Although step-wise reasoning has been proposed through textual chain-of-thought (CoT) [4,10,19,44], it remains visually blind, unable to dynamically perceive spatial misalignments [21] or evolve object states [17,27]. This motivates a paradigm shift from *language-driven* to genuinely *multimodal, visually-grounded reasoning*. However, multimodal CoT [26, 52] and tool-augmented sketching [14, 54] decouple reasoning from generation; post-hoc refinement [33] remains outcome-based, where interleaving is limited to *repairing after generation* rather than *reasoning during generation*.

We challenge this outcome-driven paradigm with **process-driven** image generation via **interleaved reasoning** anchored in both vision and text, through a unified multimodal foudnation model. We reformulate image generation with a unified multimodal models as a **co-evolving trajectory** of textual plan and visual states, orchestrated through a recurring four-stage **process**: `Plan` → `Sketch` → `Inspect` → `Refine`. The model does not hallucinate a final image; it constructs the image stroke by stroke, decision by decision. As shown in Figure 1., the first stage is `Plan`, where a unified foundation model generates text instruction `<ins>` that specifies the incremental update (what to add or modify) and the state description `<des>` of the global scene hypothesis. The `Sketch` stage synthesizes a partial visual draft conditioned on prior context. The `Inspect` stage

detects conflicts between sketch, plan, and prompt. The `Refine` stage produces a refinement plan `<refine>` and revise the image. This tightly-coupled loop enables error correction as it emerges, transforming generation from a black-box single pass into a controllable, self-correcting dialog between reasoning and vision.

A central challenge in process-driven image generation is supervising interleaved trajectories with ambiguous intermediate states, which poses three challenges: *1) how to construct a generation path and design intermediate states*: Incomplete intermediate states are inherently ambiguous. For example, a missing object could be attributed to "not yet drawn" or "incorrectly omitted". We tackle this by generating multi-turn trajectories of intermediate states via **scene graph subsampling**, yielding logically ordered incremental prompts that expand the composition without contradictions. *2) how to teach a model to see its mistakes*: We propose to construct a **dual-stream process-critique data** comprising error traces and alignment evaluation, to teach models to perform self-assessment from btoh textual and visual states, leveraging judges from a VLM operating on sampled trajectories. Finally, *how to jointly co-evolve text and vision reasoning*. We propose training a **unified multimodal sequencer**, such as BAGEL [6], to autoregressively generate *interleaved multimodal* tokens. As the result, the single model performs tasks decomposition, visual generation, self-validation and refinement in a sequential fashion, without the help from any external models.

Experiments prove our process-driven generation lifts the base BAGEL-7B [6] from 79% to 83% (+4%) on GenEval for composition object alignment and boosts the WISE for world knowledge reasoning from 70% to 76% (+6%). Compared to existing process-level approach PARM [13], our framework achieves a superior balance of performance and efficiency: it delivers higher accuracy (0.83 vs. 0.77 on GenEval) with an 8x reduction in both training data and inference cost. This advantage stems from our use of semantic partitioning—supervising concrete visual states rather than the blurry latent noise used in prior work. Finally, ablation studies identify diverse editing instructions and self-sampled critiques as the key drivers of our success, outperforming symbolic corrections by +6% and proving that internalizing the model's own failure modes is far more effective than imposing external fixes.

2 Related Work

2.1 Unified Multimodal Model

Unified multimodal models aim to unify visual understanding and generation within a single framework, building on the strong perceptual abilities of modern multimodal large language models. Early autoregressive approach, e.g., Chameleon [39], Emu3 [42], and Show-o [49], rely on discrete visual tokenizers such as VQ-VAE [29] to model images as token sequences, achieving unified model but suffering from constrained fine-grained visual understanding. Another line of work couples a pretrained LLM with an external diffusion module, where the LLM provides semantic condition for image generation [7, 31, 40, 46]; while effective, this

decoupled design prevents the model from fully leveraging its understanding capability during generation. More recent integrated transformer frameworks, such as the Janus series [3, 25], LlamaFusion [38], and BAGEL [6], directly combine autoregressive text modeling with diffusion-based generation to better align representation spaces and support large-scale interleaved text–image pretraining. However, despite these advances, existing unified models still struggle to tightly couple semantic reasoning with the generative process, limiting their ability to produce images with complex, logically structured content.

2.2 Reasoning in Image Generation

Recent studies have begun exploring interleaved reasoning in image generation, extending the success of chain-of-thought [44] from text [43, 48] to multimodal settings [5,15,30]. Early works adopt verification-based [13] or prompt-refinement strategies [22], where reasoning is performed either after image sampling or before generation; however, such text-only reasoning is isolated and remains decoupled from the evolving visual state. More recent attempts introduce multi-turn reasoning that alternates between textual analysis and visual outputs [6,9,16,51], yet these methods typically treat images as static endpoints rather than intermediate states to be interpreted, critiqued, and updated. As a result, the reasoning flow becomes fragmented and fails to maintain coherence across steps, limiting fine-grained control over spatial relations, object dynamics, and global scene evolution. Overall, current approaches do not realize fully interleaved reasoning — where textual reasoning and visual generation mutually inform each other throughout the process—highlighting a key gap our work aims to mitigate.

3 Method

Most existing image generation models generate images in a single forward pass, sometimes augmented with chain-of-thought reasoning applied exclusively to the textual prompt. However, complex spatial relationship and fine-grained visual details are inherently difficult to encode through this one-shot paradigm, as the model must resolve the entire scene before any visual feedback is available. To overcome these limitations, we introduce process-driven image generation, which reinterprets image generation as a trajectory of interleaved intermediate states. Our method alternates between textual and visual rationales over multiple steps, enabling the model to **plan**, **sketch**, **inspect**, and **refine** the scene progressively and iteratively as generation unfolds.

3.1 Framework

The general framework of our model is illustrated in Fig. 2. At a high level, our model performs image generation as a sequential, interleaved textual–visual reasoning process. Given a unified multimodal model $\mathcal{P}_\theta$ and an input text prompt T (with optional input image I_{input} for editing use case), the model generates a

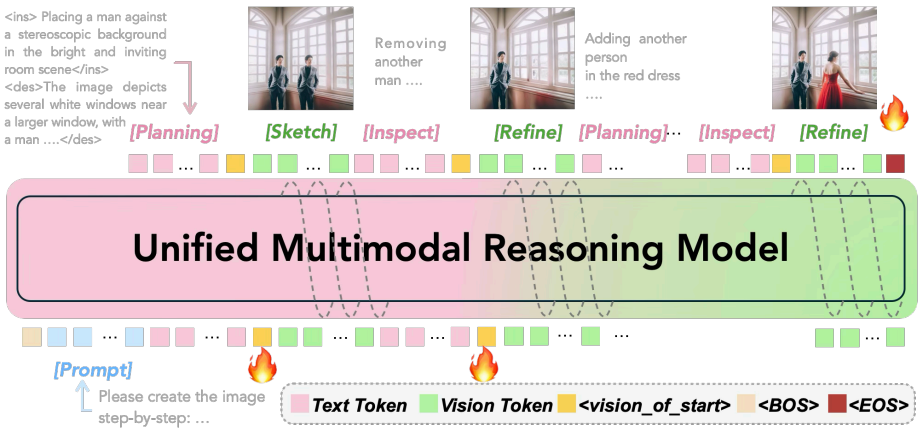

Fig. 2: We design unified multimodal reasoning models for process-driven generation, autoregressively generates an interleaved sequence of **text** and **vision tokens**.

trajectory composed of alternating textual reasoning steps $s^{(i)}$ and intermediate visual states $v^{(i)}$, ultimately converging to the final image I:

$$\{s^{(1)}, v^{(1)}, s^{(2)}, v^{(2)}, ..., v^{(k)}, s^{(k)}\}, I \sim \mathcal{P}_\theta(\cdot|T) \tag{1}$$

Building on this high-level structure, the model constructs the image through a recurring four-stage cycle: Plan, Sketch, Inspect, and Refine. Each cycle incrementally advances the generation trajectory, enabling fine-grained control over both textual and visual evolution. Concretely:

- **Stage I (Plan)**: The model interprets the prompt and accumulated context to produce an incremental instruction and a global scene description of the intended generation.
- **Stage II (Sketch)**: Conditioned on the planned instruction, the model synthesizes an updated draft image that reflects the intended modification.
- **Stage III (Inspect)**: The model inspects a) the textual incremental instruction and global scene description against the raw complete prompt and b) the produced draft against the planned textual instruction to identify potential inconsistencies or mismatches.
- **Stage IV (Refine)**: If discrepancies are found, the model revises the instruction and generates a corrected visual update, ensuring the evolving scene remains coherent and aligned with the prompt.

Through the repeated application of these four stages, the overall generation process is decomposed into a sequence of controllable, localized updates, allowing the model to progressively assemble the final image.

Each textual intermediate $s^{(i)}$ appears in two forms depending on the stage of the cycle. During the planning phase, $s^{(i)}$ includes a step-specific painting instruction enclosed in `<ins>...</ins>` and a global description enclosed in

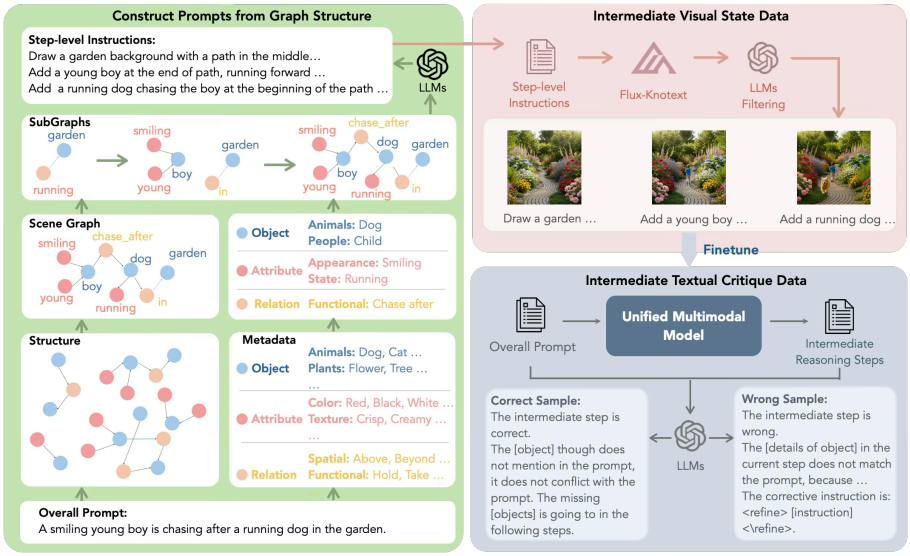

Fig. 3: Our multi-stage dataset generation pipeline constructs process interleaved trajectories with intermediate visual states and textual critiques. We ensure consistent intermediate visual state generation using scene-graph structures, and generate intermediate textual critiques via self-sampling.

`<des>...</des>`. During the inspection phase, if misalignment is detected, the model emits a refinement signal enclosed in `<refine>...</refine>`.

The corresponding visual states $v^{(i)}$ also take two forms: the planning stage produces a rough sketch that represents the intended update, while the refinement stage polishes this sketch into a more accurate visual representation. All visual outputs are wrapped between `<|vision_start|>` and `<|vision_end|>` to explicitly mark modality transitions.

Overall, the interleaved reasoning pipeline can be summarized as:

$$T \to s^{(1)}_{\text{plan}} \to v^{(1)}_{\text{sketch}} \to s^{(1)}_{\text{inspect}} \to v^{(1)}_{\text{refine}} \to \ldots \to I \tag{2}$$

3.2 Intermediate Reasoning Collection

Advancing interleaved reasoning requires addressing a challenge: representing and evaluating partially formed images. Unlike fully rendered outputs, intermediate visual states are inherently incomplete, making it difficult to judge whether the evolving scene preserves correct spatial structure and semantic coherence.

To equip the model with the ability to reason over such intermediate states, we introduce a multi-stage data construction pipeline that generates high-quality, process-oriented interleaved reasoning traces, as illustrated in Fig. 3. We then performed supervised finetuning on this dataset to teach the model how to plan,

Table 1: Detailed statistics of our curated process-oriented image-text dataset for interleaved reasoning. The dataset comprises three specialized components tailored to systematically upgrade the model's generation and reasoning capabilities: (1) *Multi-turn Generation* provides over 32K trajectories (averaging 3.51 steps per sample) to train sequential planning (s_{plan}) and step-level sketching (v_{sketch}); (2) *Instruction-Intermediate Conflict* contains ∼15K text-side samples to supervise conflict inspection (s_{inspect}); and (3) *Image-Instruction Alignment* consists of 15K balanced visual-text pairs to facilitate error analysis and state refinement (v_{refine}).

Statistic	**Number**
Multi-turn Generation	
Total Samples	32,012
Average Prompt Length	152.8
Average Image per Sample	3.51
Maximum Image per Sample	5
Instruction-Intermediate Conflict	
Total Samples	15,201
- Positive Samples	6,905
- Negative Samples	8,296
Image-Instruction Alignment	
Total Samples	15,000
- Positive Samples	5,000
- Negative Samples	10,000

assess, and refine visual states throughout the generation trajectory. Specifically, we introduced three dataset pipelines to comprehensively improve the models capability on planning, step-by-step generations and reasoning. We introduce them in detail below:

Multi-Turn Generation Subset. This dataset serves as the basic source to adapt a single-round generation model to multi-stage process-driven generation. We proposed a scene-graph based sampling mechanism to ensure at each step, only a specific local region was updated while the spatial and semantic coherence of the rest regions are preserved. Specifically, we represent each prompt using a scene graph, where object nodes, attribute nodes, and relation edges define the target composition. By subsampling subgraphs from the full scene graph, we derive a sequence of incremental step-level prompts that naturally expand the scene in a correct and controllable order — ensuring that intermediate steps remain logically grounded. We then synthesize the ground truth of each prompt using Flux-Kontext and filter with GPT.

However, subgraph expansion alone yields primarily additive updates, which limits the diversity of the prompts. To enrich the action space of visual evolution, e.g., attribute modification, swapping, removal, we further augment it by rewriting a subset of step instructions with GPT, generating semantically equivalent but structurally different multi-step reasoning variants. This increases the cov-

erage of visual editing behaviors and encourages the model to internalize richer transformation primitives beyond simple addition.

Finetuning with the Multi-Turn Generation subset provides the model with capabilities to perform step by step generation (i.e., performing the s^k_{plan} and the v^k_{sketch}), but this supervision alone is insufficient for the model to generalize to high-quality reasoning. We further introduced two reasoning subsets to boost the reasoning capabilities of the model for performing $s^k_{inspect}$ and s^k_{refine}, which teaches the model how to interpret and reason about intermediate visual states, enabling it to distinguish valid prior visual knowledge from actual conflicts with the overall prompt.

Instruction-Intermediate Conflict Reasoning Subset. This subset focuses on improving the reasoning capabilities on the textual side. To this purpose, we adopt a self-sampling strategy: from a fine-tuned model on Multi-Turn generation Subset, we sample generated intermediate reasoning traces that include textual descriptions of partially completed images and leverage GPT as judge to eval its consistency with the original raw prompt. For conflicts, we generate a textual analysis and a corrective instruction along with the reasoning. This procedure provides explicit supervision for distinguishing incomplete-but-correct intermediate textual tokens from prompt-inconsistency reasoning.

Image–Instruction Alignment Reasoning Subset. The second subset focuses on evaluating the misalignment from the visual side using the image draft and step-level painting instruction. We extend and refine Gen-Ref dataset [55] into two annotated categories: positive samples, where the image is consistent with the instruction, for which GPT generates an explanation of why the alignment holds, and negative samples, where the image mis-aligns with the instruction, for which GPT provides an error analysis and a refinement instruction.

Table 1. summarizes the key statistics of our curated dataset, which contains three major components tailored for process-driven image generation analysis. The Multi-turn Generation subset includes over 32K samples, each with 3–5 intermediate visual states on average, reflecting diverse multi-step reasoning trajectories. The Instruction-Intermediate Conflict subset provides more than 15K samples, covering both positive and negative cases to supervise the model's ability to detect and correct instruction conflicts. Additionally, the Image-Instruction Alignment subset comprises 15K samples with balanced positive and negative examples to support evaluation of fine-grained visual-text alignment.

3.3 Model

To enable process-driven interleaved reasoning, the model must possess unified multimodal understanding and capabilities. Therefore, we adopt a unified multimodal model as our backbone and finetune it for our interleaved process driven generation task.

Training Objectives. We train our model to generate text tokens autoregressively, optimizing the process with a Cross-Entropy (CE) Loss. The loss is applied to positions corresponding to textual segments $s^{(i)}$. In order to natively

generate interleaved sequences, we add a loss term on the <|vision_start|> and <|vision_end|> tokens, enabling the seamless switch between textual and visual tokens generation. The CE loss for next-token prediction is defined as,

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{text}} = -\sum_{t\in[1,i]} \log \mathcal{P}_\theta\big(s_t \mid y_{<t}, T\big) \tag{3}$$

On the visual side, we followed [6] and employ Rectified Flow paradigm [24] to generate images, following:

$$z_t^{(i)} = t \cdot z_0^{(i)} + (1-t)\cdot z_1^{(i)}, \quad t \in [0,1] \tag{4}$$

$$\mathcal{L}_{\mathrm{MSE}}^{\mathrm{image}} = \mathbb{E}\left[\left\|\mathcal{P}_\theta\big(z_t^{(i)} \mid y_{<t}, T\big) - (z_0^{(i)} - z_1^{(i)})\right\|^2\right], \tag{5}$$

The model conditions on entire preceding contexts, consisting of the input prompt and the chain of interleaved reasoning trajectories (including visual components, corresponding textual guidance, etc). The overall object for our training is a weighted combination of the above objectives.

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{CE}} \cdot \mathcal{L}_{\mathrm{CE}}^{\mathrm{text}} + \mathcal{L}_{\mathrm{MSE}}^{\mathrm{image}} \tag{6}$$

where the hyperparameter λ_{CE} is a scaling coefficient to balance CE loss and MSE loss.

Inference. During inference, given a textual prompt T, the model autoregressively generates an interleaved reasoning trajectory. Textual and visual intermediates are produced in a unified sequence, where modality shifts are governed by special tokens. The whole process terminates once the the final completed image I is emitted, marked by an end-of-sequence token.

4 Experiment

4.1 Implementation Details

We adopt the unified multimodal understanding and generation model BAGEL-7B [6] as our base model. Throughout the training, all the model parameters are finetuned end-to-end on a node with 8 NIVIDA H100 GPUs for 10,000 steps using a packed sequence of 33,000 tokens, a learning rate of 2×10^{-5}, and cosine decay. We extend the original training objective to support seamless transitions between textual reasoning and visual generation within a single autoregressive sequence. At inference time, when encounter <|vision_start|>, the model seamlessly switches to image generation mode to generate visual aids. The entire interleaved generation process only stops if the model generates the <vision_end> without <|vision_start|> following. Dataset construction details are provided in Sec. 3

Table 2: Evaluation of text-to-image generation ability on Gen-Eval benchmark. "Generation Only" stands for an image generation model, and "Unified Multimodal" denotes a model that has both understanding and generation capabilities. * means we report the reproducing results using the official Github repository and checkpoint. Our approach establishes a new state-of-the-art among unified multimodal models, delivering a 4% absolute improvement over the BAGEL base.

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall↑
Generation Only							
PixArt-Σ [2]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SDv2.1 [36]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
DALL-E 2 [35]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
Emu3-Gen [42]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SDXL [32]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3 [1]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium [8]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (12B) [18]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified Multimodal							
Chameleon [39]	–	–	–	–	–	–	0.39
LWM [23]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
SEED-X [11]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
TokenFlow-XL [34]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
ILLUME [41]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
Transfusion [53]	–	–	–	–	–	–	0.63
Emu3-Gen [42]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Janus [45]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus-Pro-7B [3]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Show-o [49]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Show-o2 [50]	1.00	0.87	0.58	0.92	0.52	0.62	0.76
BAGEL-7B* [6]	0.99	0.95	0.76	0.87	0.51	0.60	0.79
STIR (Ours)	0.99	0.95	0.75	0.87	0.72	0.69	**0.83**

4.2 Quantitative Evaluation

Table 2. demonstrates the quantitative results on the GenEval benchmark [12], which evaluates compositional text-to-image in various object-centric attributes. The evaluation includes both generation-only models and unified multimodal models. Our method exhibits particularly large gains on relational and attribute-sensitive tasks, such as position and color attribute. These categories require precise spatial reasoning and fine-grained cross-model alignment, which single-pass generative models frequently fail to capture and unified multimodal models often struggle to integrate coherently. In contrast, our process-driven interleaved reasoning yields more reliable object grounding and attribute consistency, enabling our model to match or surpass the best-performing unified systems (e.g., Janus-Pro-7B and BAGEL) while maintaining strong performance on simple single-object cases. Overall, our method sets a new state of the art among unified models, demonstrating that interleaving text–visual reasoning substantially enhances structural fidelity in image generation.

Table 3. demonstrates the quantitative results on the WISE [28] benchmark, which is designed to assess world knowledge reasoning in text-to-image generation. Generation-only models achieve moderate performance overall, ranging from 0.32 to 0.50, due to their limited multimodal understanding capabilities.

Table 3: Evaluation of world knowledge reasoning WISE benchmark. WISE assesses a model's ability to integrate world knowledge and structured semantic reasoning into text-to-image generation. Our approach boosts BAGEL by 6% absolute gains, achieving nearly 15% gains on challenging tasks like Time and Chemistry.

Model	Culture	Time	Space	Biology	Physics	Chemistry	Overall↑
Generation Only							
SDv1.5 [37]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SDXL [32]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD3.5-large [8]	0.44	0.50	0.58	0.44	0.52	0.31	0.44
PixArt-Σ [2]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Playground-v2.5 [20]	0.49	0.58	0.55	0.43	0.48	0.33	0.49
FLUX.1-dev [18]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified Multimodal							
Janus [45]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
VILA-U [47]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Show-o [49]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Janus-Pro-7B [3]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3 [42]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
MetaQuery [31]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
Show-o2 [50]	0.64	0.58	0.61	0.58	0.63	0.49	0.61
BAGEL [6]	0.76	0.69	0.75	0.64	0.75	0.58	0.70
STIR (Ours)	0.74	0.82	0.73	0.70	0.76	0.78	**0.76**

Unified multimodal models, such as Janus-Pro and BAGEL, exhibit stronger results. However, these models struggle with temporal and scientific domains, such as Time and Chemistry. By incorporating process-driven reasoning, our model achieves the best overall score 0.76 and delivers consistent improvements across nearly all domains. In particular, we observe substantial gains in Time, Biology, and Chemistry, demonstrating enhanced generalization to complex concepts. These results show that interleaved reasoning trajectories enable models to better utilize world knowledge during generation.

4.3 Analysis of Process-driven Reasoning

To evaluate the effectiveness of our proposed paradigm, we conduct a comparative analysis against several process-based baselines. This experiment aims to verify whether our model effectively internalizes visual-semantic grounding and to demonstrate its efficiency compared to existing training-free and training-based workflows.

We evaluate our approach against two distinct categories of baselines:

- **BAGEL + GPT (Planner):** Use GPT-4o as external guidance to provide step-by-step instructions for multi-turn image generation by the vanilla BAGEL model.
- **BAGEL + GPT (Inspector):** Employ BAGEL to perform a single-round refinement of its initial draft conditioned on verbal feedback regarding inconsistencies provided by a GPT-4o.

Table 4: Qualitative comparison with process-driven baselines on Gen-Eval benchmark. "Training Dataset" indicates the number of samples used for fine-tuning. "Inference Cost" denotes the cumulative of sampling steps required to synthesize a final image. Our method achieves superior performance with significantly lower training and inference overhead compared to existing process-based approaches.

Reasoning Strategy	Training Dataset	Inference Cost	Gen-Eval Bench
Training-free			
BAGEL [6] + GPT (Planner)	-	50	0.60
BAGEL [6] + GPT (Inspector)	-	50	0.80
Training-based			
PARM [13] (TTS)	400K	1000	0.67
PARM [13] (RL+TTS)	688K	1000	0.77
STIR (Ours, SFT)	62K	131	0.83

– **PARM:** A state-of-the-art baseline that performs step-level verification by assessing the potential of intermediate, blurry decoding states to satisfy the prompt, used for path selection (TTS) or iterative alignment (RL).

The comparison with training-free methods in Table 4. underscores that **SFT is indispensable for internalizing visual-semantic grounding.** Attempting to use GPT-4o as an external planner for the base model leads to a 23% performance collapse. This occurs because, without specialized tuning, the base model cannot consistently ground incremental, multi-step instructions into stable visual intermediates. Furthermore, external verbal critiques from an inspector provide only marginal gains, confirming that verbal feedback cannot be effectively translated into precise visual updates unless the model's native generative path is aligned through process-level training. Our approach bridges this gap, enabling a 7B model to surpass the performance of the 12B FLUX.1-dev and even surpass it by 26% on the WISE benchmark.

Compared to training-based baselines PARM, our method demonstrates a superior balance between performance and efficiency across multiple dimensions:

– **Computational and Data Efficiency:** ur model establishes a new SOTA score of 0.83 on GenEval, outperforming even the RL-enhanced PARM (0.77). Crucially, we achieve this using only **62K** samples—an 11× reduction in training data compared to PARM's 688K. During inference, our model provides a nearly 8× speedup. While PARM relies on a Best-of-20 search strategy that incurs a cumulative cost of 1000 sampling steps, our model synthesizes high-quality images in just 131 steps (averaging 2.62 reasoning steps per image). This inference cost is calculated based on our model's complexity-adaptive reasoning, where it autonomously determines the trajectory length based on task difficulty.
– **Visual-Semantic Partitioning vs. Pixel-level Supervision:** A fundamental distinction lies in the nature of the "process" itself. PARM supervises the diffusion latent space, where intermediate states are often fuzzy or lack

Table 5: Diverse editing instructions unlock relational reasoning: augmenting additive step prompts with richer operations boosts performance.

Case	Color	Position	Color Attr.
w/o aug.	0.81	0.58	0.50
+ Self-critique	0.84	0.61	0.53
w/ aug.	0.82	0.67	0.62
+ Self-critique	0.87	0.72	0.69

Table 6: Supervising refinement via the model's own error trajectories (*self-sampling*) yields better performance over scene-graph-derived corrections.

Case	Color	Position	Color Attr.
	0.82	0.67	0.62
+ scene graph	0.83	0.70	0.67
+ self-sampling	0.87	0.72	0.69

clear human-interpretable meaning. In contrast, we utilize semantic partitioning, where reasoning trajectories are adaptively driven by the emergence of objects and relations. By supervising the model on trajectories with concrete visual semantics rather than abstract noise levels, our model achieves superior grounding and complexity-adaptive reasoning.

4.4 Ablation Study and Analysis

To understand the contribution of each component in our process-driven recipe, we conduct an extensive ablation study on the GenEval benchmark across three dimensions: the structure of step-level instructions, the construction of intermediate supervision, and the design of semantic–visual consistency checks. The results show that intermediate reasoning quality emerges from the interaction of these components rather than any single factor.

Diverse step-level instructions enable model to develop more flexible intermediate reasoning strategies. To investigate how the form of step-level instructions influences multi-turn generation, we compare two variants of our multi-turn generation dataset. The first variant constructs instructions from incrementally expanded subgraphs of the scene graph, resulting in a sequence of purely additive editing operations. The second variant augments this dataset by rewriting a subset of instructions using ChatGPT, introducing a richer set of editing types such as attribution modification, swapping, removal, etc.

Table 5. shows that while additive-only supervision (*w/o aug.*) yields moderate results, introducing instruction diversity (*w/ aug.*) drives substantial gains, specifically +0.09 in positional and +0.12 in attribute accuracy. This advantage becomes more pronounced after self-critique fine-tuning, with the performance gap in positional grounding widening to +0.11. This suggests that intermediate process supervision benefits from exposing the model to varied forms of visual reasoning, enabling it to better interpret evolving visual states and execute more flexible and coherent edits, rather than following a single monotonic editing trajectory. By internalizing non-linear transformations like swapping or removal, the model develops a flexible reasoning behavior that mirrors real-world creative workflows, allowing it to interpret partially formed images as adjustable states requiring active refinement.

Table 7: Complementary designs on intermediates states improve different generation tasks: Instruction-intermediate conflict supervision (w/ ins.) primarily improves semantic and spatial consistency, while Image-Instruction alignment supervision (w/ img-ins.) yields gains in visually grounding.

Case	Counting	Colors	Position	Color Attr.
Multi-turn SFT	0.61	0.84	0.66	0.62
+ Instruction-intermediate conflict (w/ ins.)	0.62	0.85	0.71	0.65
+ mage-Instruction alignment (w/ img-ins.)	0.73	0.86	0.69	0.65
w/ ins. + img-ins. (ours)	0.75	0.87	0.72	0.69

The consistency of critique space is more important than the controability. To investigate the optimal construction of intermediate critique supervision, we compare two data generation strategies: Symbolic Corrections, which derives corrections from the scene graph: we target a specific object, attribute, or relation and produce a corresponding critique, and Self-Sampling, where GPT identifies alignment conflicts within model-generated visual trajectories and proposes corrective instructions

As shown in Table 6., while symbolic edits offer marginal gains, self-sampling drives significant improvements over the multi-turn baseline, raising Color (0.87 vs. 0.82), Position (0.72 vs. 0.67), and Attribute (0.69 vs. 0.62) accuracy by substantial margins. We attribute this improvement to the fact that self-sampling operates in the model's own distribution. The critique data reflects the model's actual failure modes and correction needs, enabling the supervision signal to be aligned with the model's internal reasoning dynamics.

Semantic and visual constraints address distinct failure modes and jointly improve compositional reasoning accuracy. Table 7 demonstrates that our two supervision mechanisms are complementary: instruction-intermediate conflict checks (*w/ ins.*) mitigate instruction drift and improve positional accuracy by +5%, while image-instruction alignment (*w/ img-ins.*) sharpens grounded reasoning, yielding a +12% boost in Counting tasks. Their combination achieves peak performance across all metrics, confirming that enforcing correctness during intermediate steps—rather than solely at the final output—is essential to detect and correct inconsistencies before they propagate through reasoning.

4.5 Qualitative Evaluation

Fig. 4. illustrates the reasoning trajectories produced by our process-driven generation paradigm. The first row shows that the model transforms conventional single-pass generation into a sequence of adaptive reasoning steps, progressively refining both the textual plan and the visual draft. The second and third rows highlight the model's ability to detect and correct two different types of intermediate errors: (1) conflicts between step-level instruction and overall prompt, and (2) mismatches between instruction and partially generated image. These complementary mechanisms enable the model to revise wrong intermediate states

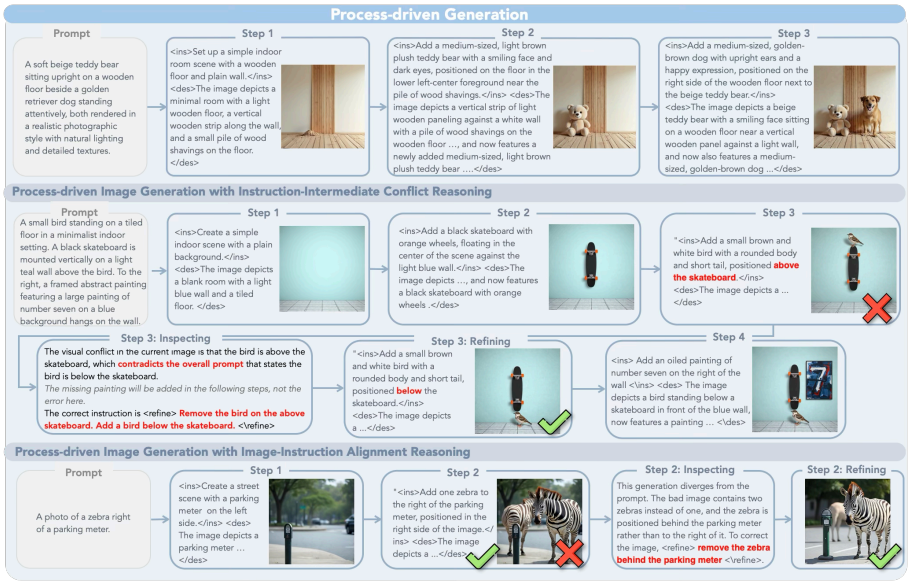

Fig. 4: Visualization of the interleaved reasoning trajectory in our process-driven image generation. Each step follows the plan–sketch–inspect–refine cycle, while `inspect` steps with no detected issues are omitted for brevity. The second and third rows illustrate two types of intermediate errors: (1) conflicts between the step-level instruction and the overall prompt, where the model revises the instruction and corrects the image; and (2) inconsistencies between the generated draft and instruction, where the instruction and overall prompt remain valid but the image requires refinement.

and provide a coherent context for subsequent updates. Notably, the model does not misinterpret incomplete or yet-to-be-rendered details as errors, demonstrating its capacity to distinguish intermediate progress from true inconsistencies.

5 Conclusion

We introduce a novel process-driven interleaved reasoning paradigm that teaches a unified multimodal model to build images stroke by stroke, decision by decision, via a co-evolving loop of textual planning, visual sketching, self-inspection, and refinement. Our method hinges on three breakthroughs: scene-graph subsampling for contradiction-free incremental instructions, self-sampled critique traces to learn from the model's errors, and end-to-end training of BAGEL to autoregressively emit interleaved text and image tokens. We lift the public BAGEL from 0.79 to 0.83 on GenEval and from 0.70 to 0.76 on WISE. Looking forward, we will extend unified multimodal reasoning to videos and 3D space, and enable real-time human-in-the-loop control, unlocks controllable, truthful and interpretable image synthesis.

References

1. Betker, J., Goh, G., Jing, L., TimBrooks, Wang, J., Li, L., LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, Ramesh, A.: Improving image generation with better captions. https://api.semanticscholar.org/CorpusID:264403242
2. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation (2024)
3. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling (2025), https://arxiv.org/abs/2501.17811
4. Creswell, A., Shanahan, M., Higgins, I.: Selection-inference: Exploiting large language models for interpretable logical reasoning (2022), https://arxiv.org/abs/2205.09712
5. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., et al., A.L.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: Dreamllm: Synergistic multimodal comprehension and creation (2024), https://arxiv.org/abs/2309.11499
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://arxiv.org/abs/2403.03206
9. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., Liu, X., Li, H.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing (2025), https://arxiv.org/abs/2503.10639
10. Feng, Y., Chen, X., Lin, B.Y., Wang, P., Yan, J., Ren, X.: Scalable multi-hop relational reasoning for knowledge-aware question answering. In: EMNLP. pp. 1295–1309 (Nov 2020)
11. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation (2025), https://arxiv.org/abs/2404.14396
12. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment (2023), https://arxiv.org/abs/2310.11513
13. Guo, Z., Zhang, R., Tong, C., Zhao, Z., Huang, R., Zhang, H., Zhang, M., Liu, J., Zhang, S., Gao, P., Li, H., Heng, P.A.: Can we generate images with cot? let's verify and reinforce image generation step by step (2025), https://arxiv.org/abs/2501.13926
14. Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models (2024), https://arxiv.org/abs/2406.09403
15. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models (2025), https://arxiv.org/abs/2503.06749

16. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot (2025), https://arxiv.org/abs/2505.00703
17. Jung, J.H., Kim, E.T., Kim, S., Lee, J.H., Kim, B., Chang, B.: Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning (2025), https://arxiv.org/abs/2411.16761
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
19. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought (2025), https://arxiv.org/abs/2501.07542
20. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation (2024), https://arxiv.org/abs/2402.17245
21. Li, L., Bigverdi, M., Gu, J., Ma, Z., Yang, Y., Li, Z., Choi, Y., Krishna, R.: Unfolding spatial cognition: Evaluating multimodal models on visual simulations (2025), https://arxiv.org/abs/2506.04633
22. Liao, J., Yang, Z., Li, L., Li, D., Lin, K., Cheng, Y., Wang, L.: Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning (2025), https://arxiv.org/abs/2503.19312
23. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with blockwise ringattention (2025), https://arxiv.org/abs/2402.08268
24. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
25. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., yu, X., Zhao, L., Wang, Y., Liu, J., Ruan, C.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation (2025), https://arxiv.org/abs/2411.07975
26. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain of thought prompting for large multimodal models. In: CVPR (2024)
27. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought (2023), https://arxiv.org/abs/2305.15021
28. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation (2025), https://arxiv.org/abs/2503.07265
29. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018), https://arxiv.org/abs/1711.00937
30. OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A.T., Neitz, A., Prokofiev, A., et al., A.W.: Openai o1 system card (2024), https://arxiv.org/abs/2412.16720
31. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., Hou, J., Xie, S.: Transfer between modalities with metaqueries (2025), https://arxiv.org/abs/2504.06256

32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023), https://arxiv.org/abs/2307.01952
33. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision (2025), https://arxiv.org/abs/2508.05606
34. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation (2025), https://arxiv.org/abs/2412.03069
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022), https://arxiv.org/abs/2204.06125
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
38. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Lmfusion: Adapting pretrained language models for multimodal generation (2025), https://arxiv.org/abs/2412.15188
39. Team, C.: Chameleon: Mixed-modal early-fusion foundation models (2025), https://arxiv.org/abs/2405.09818
40. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning (2024), https://arxiv.org/abs/2412.14164
41. Wang, C., Lu, G., Yang, J., Huang, R., Han, J., Hou, L., Zhang, W., Xu, H.: Illume: Illuminating your llms to see, draw, and self-enhance (2024), https://arxiv.org/abs/2412.06673
42. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., Zhao, Y., Ao, Y., Min, X., Li, T., Wu, B., Zhao, B., Zhang, B., Wang, L., Liu, G., He, Z., Yang, X., Liu, J., Lin, Y., Huang, T., Wang, Z.: Emu3: Next-token prediction is all you need (2024), https://arxiv.org/abs/2409.18869
43. Wang, X., Zhou, D.: Chain-of-thought reasoning without prompting (2024), https://arxiv.org/abs/2402.10200
44. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
45. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., Luo, P.: Janus: Decoupling visual encoding for unified multimodal understanding and generation (2024), https://arxiv.org/abs/2410.13848
46. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm (2024), https://arxiv.org/abs/2309.05519
47. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., Han, S., Lu, Y.: Vila-u: a unified foundation model integrating visual understanding and generation (2025), https://arxiv.org/abs/2409.04429
48. Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., Wang, Y.: When more is less: Understanding chain-of-thought length in llms (2025), https://arxiv.org/abs/2502.07266

49. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation (2025), https://arxiv.org/abs/2408.12528
50. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models (2025), https://arxiv.org/abs/2506.15564
51. Zhang, Y., Li, Y., Yang, Y., Wang, R., Yang, Y., Qi, D., Bao, J., Chen, D., Luo, C., Qiu, L.: Reasongen-r1: Cot for autoregressive image generation models through sft and rl (2025), https://arxiv.org/abs/2505.24875
52. Zheng, G., Yang, B., Tang, J., Zhou, H.Y., Yang, S.: Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In: NeurIPS2023 (2023)
53. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model (2024), https://arxiv.org/abs/2408.11039
54. Zhou, Q., Zhou, R., Hu, Z., Lu, P., Gao, S., Zhang, Y.: Image-of-thought prompting for visual reasoning refinement in multimodal large language models (2024), https://arxiv.org/abs/2405.13872
55. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In: ICCV (2025)