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

Lei Zhang[1,2,*], **Jujian Tian**[1], **Zhipeng Fan**[1], **Kunpeng Li**[1], **Jialiang Wang**[1], **Weifeng Chen**[1], **Markos Georgopoulos**[1], **Felix Juefei-Xu**[1], **Yuxiang Bao**[3], **Julian McAuley**[2,†], **Manling Li**[4,†], **Zecheng He**[1,†]

[1]Meta Superintelligence Labs, [2]University of California, San Diego, [3]Worcester Polytechnic Institute, [4]Northwestern University

[*]Work done at Meta, [†]Joint last author

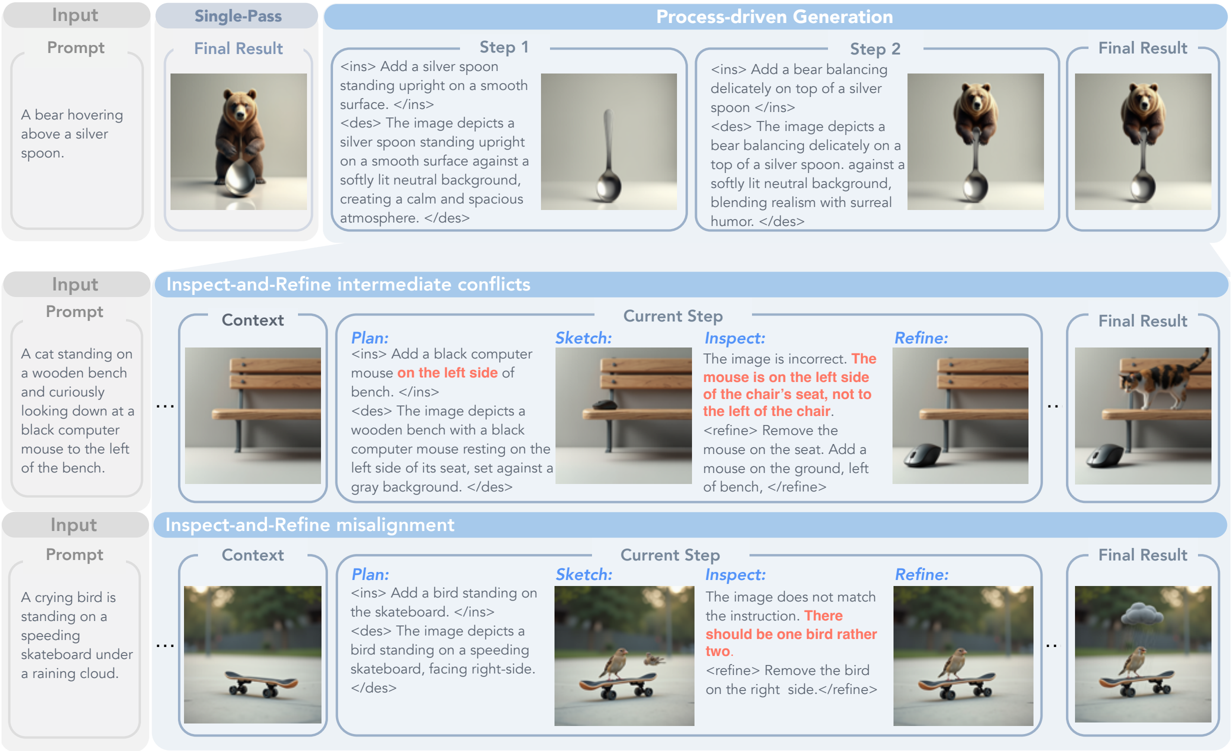
Humans paint images incrementally: they plan a global layout, sketch a coarse draft, inspect, and refine details, and most importantly, each step is grounded in the evolving visual states. However, can unified multimodal models trained on text-image interleaved datasets also imagine the chain of intermediate 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 generation benchmarks.

Date: February 11, 2026

1 Introduction

Despite the impressive progress of image generation, today's models still remain brittle on elementary visual logic and produce plausible but incorrect images. As shown in Figure 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 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) Wei et al. (2023); Creswell et al. (2022); Feng et al. (2020); Li et al. (2025a), but it remains visually blind, unable to dynamically perceive spatial misalignments Li et al. (2025b) or evolve object states Jung et al. (2025); Mu et al. (2023). This motivates a paradigm shift from *language-driven* to genuinely *multimodal, visually-grounded reasoning*. However, multimodal CoT Mitra et al. (2024); Zheng et al. (2023) and tool-augmented sketching Hu et al. (2024); Zhou et al. (2024b) decouple reasoning from generation; post-hoc refinement Qin et al. (2025) 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



A bear hovering above a silver spoon.

<ins> Add a silver spoon standing upright on a smooth surface. </ins>
<des> The image depicts a silver spoon standing upright on a smooth surface against a softly lit neutral background, creating a calm and spacious atmosphere. </des>

<ins> Add a bear balancing delicately on top of a silver spoon </ins>
<des> The image depicts a bear balancing delicately on a top of a silver spoon. against a softly lit neutral background, blending realism with surreal humor. </des>

A cat standing on a wooden bench and curiously looking down at a black computer mouse to the left of the bench.

Plan: <ins> Add a black computer mouse on the left side of bench. </ins>
<des> The image depicts a wooden bench with a black computer mouse resting on the left side of its seat, set against a gray background. </des>

Inspect: The image is incorrect. **The mouse is on the left side of the chair's seat, not to the left of the chair**.
<refine> Remove the mouse on the seat. Add a mouse on the ground, left of bench, </refine>

A crying bird is standing on a speeding skateboard under a raining cloud.

Plan: <ins> Add a bird standing on the skateboard. </ins>
<des> The image depicts a bird standing on a speeding skateboard, facing right-side. </des>

Inspect: The image does not match the instruction. **There should be one bird rather two**.
<refine> Remove the bird on the right side.</refine>

Figure 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.

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 dialogue 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 to train a **unified multimodal sequencer**, such as BAGEL Deng et al. (2025), to autoregressively generate *interleaved multi-modal* 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 on GenEval prove our process-driven generation lifts the base BAGEL-7B Deng et al. (2025) from 77% to 82% (+5%). Ablation reveals why process supervision works: diverse editing instructions (*add/refine/remove*) unlock +9% on relational tasks, and self-sampled critiques, tuned to the model's own failure modes, outperform symbolic corrections by +6%, proving that teaching a model to spot its own mistakes works better than forcing external fixes. Emergent abilities also arise with no supervision: the model automatically enriches information during reasoning, and adjusts its planning with only 1.2 steps for a single cat, but 4.5 steps for counting tasks with more objects.

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 Team (2025), Emu3 Wang et al. (2024b), and Show-o Xie et al. (2025a), rely on discrete visual tokenizers such as VQ-VAE van den Oord et al. (2018) 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 Dong et al. (2024); Wu et al. (2024b); Pan et al. (2025); Tong et al. (2024); 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 Ma et al. (2025); Chen et al. (2025), LlamaFusion Shi et al. (2025), and BAGEL Deng et al. (2025), 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 Wei et al. (2023) from text domains Wang and Zhou (2024); Wu et al. (2025b) to multimodal settings OpenAI et al. (2024); Huang et al. (2025); DeepSeek-AI et al. (2025). Early works adopt verification-based Guo et al. (2025) or prompt-refinement strategies Liao et al. (2025), 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 Fang et al. (2025); Deng et al. (2025); Jiang et al. (2025); Zhang et al. (2025), 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 Figure 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 trajectory composed of alternating textual reasoning steps $s^{(i)}$ and intermediate visual states $v^{(i)}$, ultimately converging to the final image I:

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

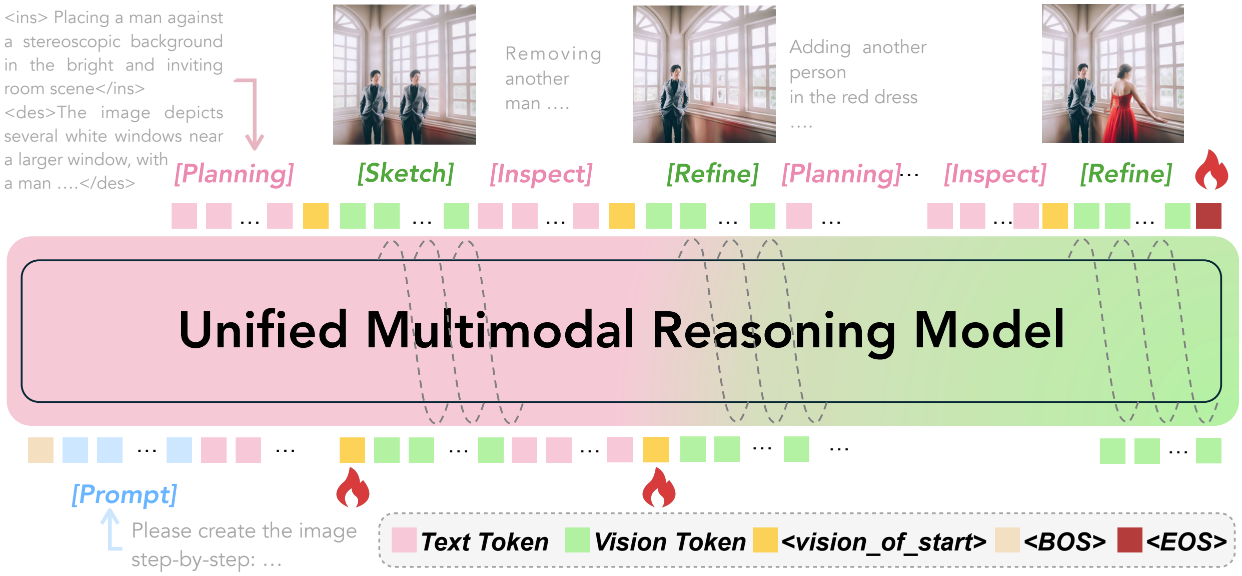

<ins> Placing a man against a stereoscopic background in the bright and inviting room scene</ins> <des>The image depicts several white windows near a larger window, with a man ....</des>
Removing another man ....
Adding another person in the red dress ....
[Planning]
[Sketch]
[Inspect]
[Refine]
[Planning]...
[Inspect]
[Refine]
Unified Multimodal Reasoning Model
[Prompt]
Please create the image step-by-step: ...
Text Token
Vision Token
<vision_of_start>
<BOS>
<EOS>

Figure 2 We design unified multimodal reasoning models for process-driven generation, autoregressively generates an interleaved sequence of **text tokens** (pink) and **vision tokens** (green).

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 `<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:

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

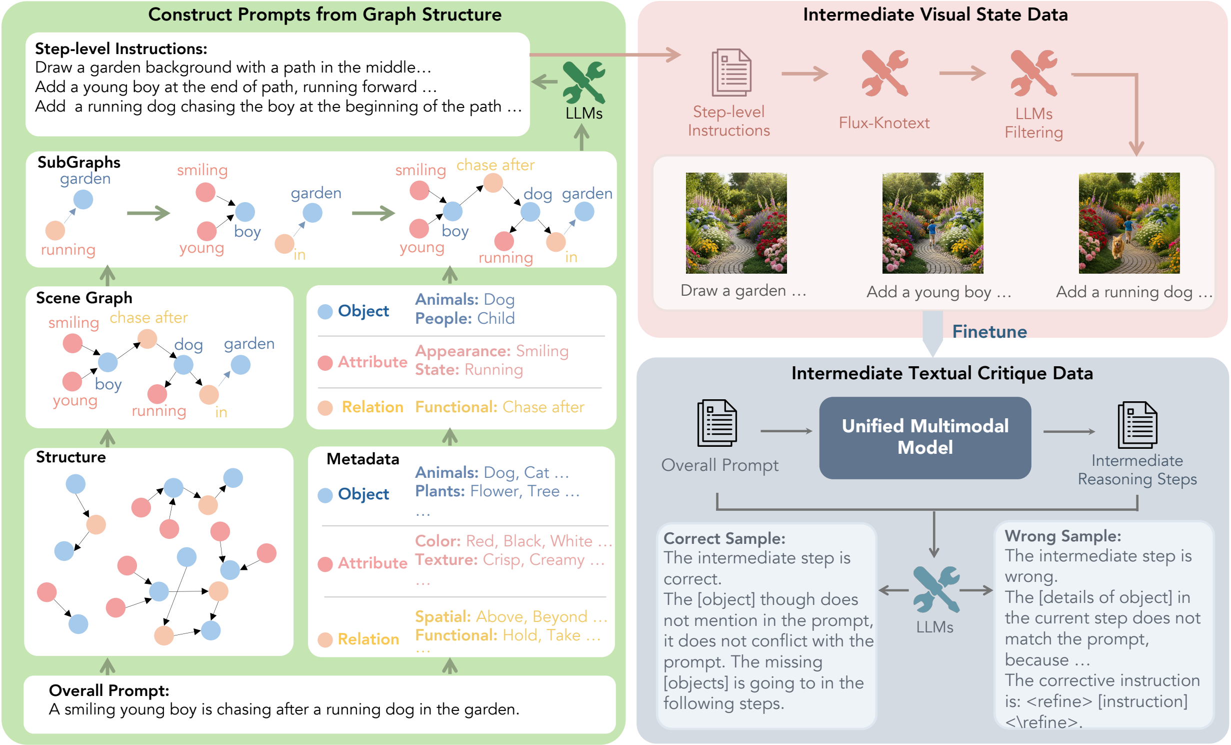

Construct Prompts from Graph Structure
Step-level Instructions:
Draw a garden background with a path in the middle…
Add a young boy at the end of path, running forward …
Add a running dog chasing the boy at the beginning of the path …
LLMs
SubGraphs
garden
running
smiling
young
boy
garden
in
smiling
chase after
young
boy
running
dog
garden
in
Scene Graph
smiling
chase after
dog
garden
boy
young
running
in
Object
Animals: Dog
People: Child
Attribute
Appearance: Smiling
State: Running
Relation
Functional: Chase after
Structure
Metadata
Object
Animals: Dog, Cat …
Plants: Flower, Tree …
…
Attribute
Color: Red, Black, White …
Texture: Crisp, Creamy …
…
Relation
Spatial: Above, Beyond …
Functional: Hold, Take …
…
Overall Prompt:
A smiling young boy is chasing after a running dog in the garden.
Intermediate Visual State Data
Step-level Instructions
Flux-Knotext
LLMs Filtering
Draw a garden …
Add a young boy …
Add a running dog …
Finetune
Intermediate Textual Critique Data
Overall Prompt
Unified Multimodal Model
Intermediate Reasoning Steps
LLMs
Correct Sample:
The intermediate step is correct.
The [object] though does not mention in the prompt, it does not conflict with the prompt. The missing [objects] is going to in the following steps.
Wrong Sample:
The intermediate step is wrong.
The [details of object] in the current step does not match the prompt, because …
The corrective instruction is: <refine> [instruction] <\refine>.

Figure 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.

3.2 Intermediate Reasoning Collection

Advancing interleaved reasoning requires addressing a core 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 Figure 3. We then performed supervised finetuning on this dataset to teach the model how to plan, 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 coverage 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 current image draft and the step-level painting instruction. We extend and refine the Gen-Ref dataset Zhuo et al. (2025) 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 both 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.

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

Table 1 Statistics of intermediate reasoning dataset.

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, such as BAGEL Deng et al. (2025), 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 only 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 formally 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 Deng et al. (2025) and employ Rectified Flow paradigm Liu et al. (2022) to generate images, following:

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

Model	Single Object	Two Objects	Counting	Colors	Position	Color Attributes	Overall↑
Generation Only							
PixArt-Σ Chen et al. (2024)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SDv2.1 Rombach et al. (2022b)	0.98	0.51	0.44	0.85	0.07	0.17	0.50
DALL-E 2 Ramesh et al. (2022)	0.94	0.66	0.49	0.77	0.10	0.19	0.52
Emu3-Gen Wang et al. (2024b)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SDXL Podell et al. (2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3 Betker et al.	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium Esser et al. (2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (12B) Labs et al. (2025)	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified Multimodal							
Chameleon Team (2025)	–	–	–	–	–	–	0.39
LWM Liu et al. (2025)	0.93	0.41	0.46	0.79	0.09	0.15	0.47
SEED-X Ge et al. (2025)	0.97	0.58	0.26	0.80	0.19	0.14	0.49
TokenFlow-XL Qu et al. (2025)	0.95	0.60	0.41	0.81	0.16	0.24	0.55
ILLUME Wang et al. (2024a)	0.99	0.86	0.45	0.71	0.39	0.28	0.61
Transfusion Zhou et al. (2024a)	–	–	–	–	–	–	0.63
Emu3-Gen Wang et al. (2024b)	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Janus Wu et al. (2024a)	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus-Pro-7B Chen et al. (2025)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Show-o Xie et al. (2025a)	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Show-o2 Xie et al. (2025b)	1.00	0.87	0.58	0.92	0.52	0.62	0.76
BAGEL-7B* Deng et al. (2025)	0.99	0.95	0.76	0.87	0.51	0.56	0.77
Ours (BAGEL-7B + Process-Driven)	0.99	0.94	0.75	0.87	0.72	0.63	**0.82**

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 boosts BAGEL-7B by 5% absolute gains, and the performance with 7B parameters achieve comparable with a 12B state-of-the-art generation-only model FLUX.1-dev.

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

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.

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

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

Dataset. We construct a process-based interleaved reasoning dataset from scratch, consisting of three complementary components: the multi-turn generation subset, instruction-intermediate conflict set and image-instruction alignment set. The multi-turn generation subset contains 30K samples, each paired with approximately three intermediate images and corresponding step-level instructions. To support prompt–intermediate conflict reasoning, we further collect 15K samples by self-sampling intermediate trajectories as the instruction-intermediate conflict set. Finally, the image–instruction alignment subset includes 15K image–text pairs (10K negative and 5K positive) to enforce fine-grained consistency between editing instructions and resulting visual modifications. Additional dataset construction details are provided in Sec. 3.

Model	Culture	Time	Space	Biology	Physics	Chemistry	Overall↑
Generation Only							
SDv1.5 Rombach et al. (2022a)	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SDXL Podell et al. (2023)	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD3.5-large Esser et al. (2024)	0.44	0.50	0.58	0.44	0.52	0.31	0.44
PixArt-Σ Chen et al. (2024)	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Playground-v2.5 Li et al. (2024)	0.49	0.58	0.55	0.43	0.48	0.33	0.49
FLUX.1-dev Labs et al. (2025)	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified Multimodal							
Janus Wu et al. (2024a)	0.16	0.26	0.35	0.28	0.30	0.14	0.23
VILA-U Wu et al. (2025a)	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Show-o Xie et al. (2025a)	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Janus-Pro-7B Chen et al. (2025)	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3 Wang et al. (2024b)	0.34	0.45	0.48	0.41	0.45	0.27	0.39
MetaQuery Pan et al. (2025)	0.56	0.55	0.62	0.49	0.63	0.41	0.55
Show-o2 Xie et al. (2025b)	0.64	0.58	0.61	0.58	0.63	0.49	0.61
BAGEL Deng et al. (2025)	0.76	0.69	0.75	0.64	0.75	0.58	0.70
Ours (BAGEL + Process-driven)	0.74	0.82	0.73	0.70	0.76	0.78	**0.76**

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. "Generation Only" stands for an image genreation model, and "Unified Multimodal" models jointly support both understanding and generation. Our approach boosts BAGEL-7B by 8.5% absolute gains, achieving nearly 15% gains on challenging tasks like Time and Chemistry.

Training and Inference Details. We adopt the unified multimodal understanding and generation model BAGEL-7B Deng et al. (2025) 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.

4.2 Quantitative Evaluation

Table 2. demonstrates the quantitative results on the GenEval benchmark Ghosh et al. (2023), 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 Niu et al. (2025) 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. 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 Emergent Properties

Beyond performance gains, process-driven generation reveals several emergent behaviors, including information-aware reasoning, adaptive depth, and intermediate self-correction, as demonstrated in the following experiments.

Emergent Property ①: Information-Enriched Utilization The model benefits from leveraging detailed prompts, unfolding a information-enriched reasoning trajectory, while single-pass generation remains insensitive to such enrichment.

Model	Case	Two Obj.	Counting	Position
Bagel	w/o rewrite	0.95	0.76	0.50
	w/ rewrite	0.92	0.75	0.54
Ours	w/o rewrite	0.88	0.66	0.61
	w/ rewrite	0.94	0.75	0.72

Table 4 Information-enriched reasoning via prompt rewriting: Step-level instructions with richer semantic details improves compositional accuracy across key GenEval tasks.

Table 4. reveals that prompt rewriting supplies richer semantic cues that significantly benefit our process-driven generation. While Bagel exhibits limited or even negative gains from rewritten prompts—indicating that single-pass generation is relatively insensitive to additional semantic structure—our model consistently improves across all categories. Notably, the largest benefits appear in Position and Counting, suggesting that enriched prompts particularly reinforce relational and quantitative reasoning. These improvements arise because the interleaved reasoning framework explicitly consumes detailed step-level instructions throughout the generation process, enabling the model to propagate and preserve enriched semantics across multiple intermediate states. Taken together, these observations highlight an emergent property of our approach: the capability to transform information-enriched prompts into more accurate reasoning trajectories and more reliable compositional image generation.

Emergent Property ②: Complexity-Adaptive Reasoning The model shows plasticity in the depth and granularity of reasoning trajectory based on the complexity of the prompt.

As shown in Figure 4., the depth of reasoning required increases as the complexity of the task and the relational dependency increase. Single Object and Color typically require only 1-2 steps to complete, while more demanding tasks, such as Counting and Position, induce significantly long trajectories (up to 4-5 steps), where precise relational grounding is essential. Remarkably, this behavior emerges without any supervision on the number reasoning steps: the model autonomously allocates more iterative updates when the prompt involves complex interactions among objects. These findings suggest that interleaved reasoning enables a form of computational adaptivity, allowing the model to flexibly balance accuracy and efficiency by selectively deepening the reasoning process when necessary.

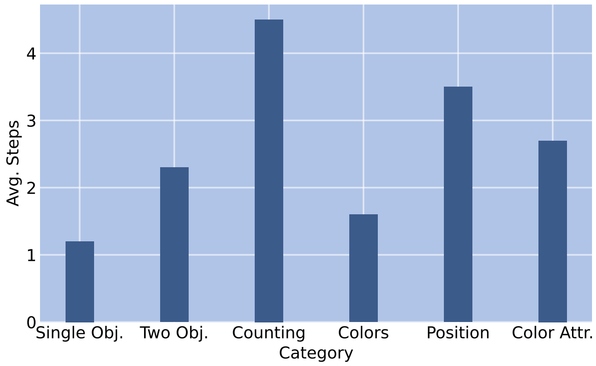


Figure 4 Complexity-adaptive reasoning emerges: The model autonomously scales inference depth with task difficulty.

4.4 Ablation Study

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.

Ablation Study ①: The Impact of Instruction Diversity Diverse step-level instructions enable model to develop more flexible intermediate reasoning strategies, leading to better performance.

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 strictly 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, and etc.

Case	Color	Position	Color Attri.
w/o aug.	0.81	0.58	0.50
+ Self-critique	0.84	0.61	0.53
w/ aug.	0.82	0.67	0.59
+ Self-critique	0.87	0.72	0.63

Table 5 Diverse editing instructions unlock relational reasoning: augmenting additive step prompts with richer operations (*refine, remove, swap*) boosts performance.

As shown in Table 5., the dataset with only additive operations (w/o aug.) achieves moderate performance consistently. Introducing instruction diversity (w/ aug.) leads to clear improvements, especially in position (+0.09) and attribute accuracy (+0.09). After applying self-critique fine-tuning, both settings improve, but the diversified-instruction variant again shows a larger boost (+0.11 in position, +0.10 in attributes), demonstrating that models benefit from learning richer forms of intermediate visual reasoning.

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. We attribute the improvement to the role of instruction diversity in shaping the model's intermediate reasoning behavior, which mirrors real-world creative workflows. Such diversity encourages the model to interpret partially formed images not simply as "incomplete additions", but as editable states that may require correction or adjustment, which also works as the basis of further inspecting and refining.

Ablation Study ②: Consistency in Critique The consistency of critique space is more important than the controability, so we use the self-sampling.

To study how intermediate critique supervision should be constructed, we compare two data generation strategies. The first strategy derives corrections directly from the scene graph: we target a specific object, attribute, or relation and produce a corresponding critique. The second strategy instead performs self-sampling: we generate intermediate visual states using a model fine-tuned on the multi-turn dataset, and then use GPT to determine whether each state aligns with the overall prompt and propose corrections when conflicts arise.

Case	Color	Position	Color Attri.
	0.82	0.67	0.59
+ scene graph	0.83	0.70	0.61
+ self-sampling	0.87	0.72	0.63

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

Table 6. compares different strategies for constructing intermediate critique supervision. Starting from the baseline multi-turn model (Color: 0.82, Position: 0.67, Attribute: 0.59), incorporating scene graph–based critiques leads to moderate improvements (0.83 / 0.70 / 0.61), suggesting that explicitly correcting individual objects, attributes, or relations can help guide the refinement process. However, replacing these symbolic edits with self-sampled critiques yields substantially larger gains (0.87 / 0.72 / 0.63).

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.

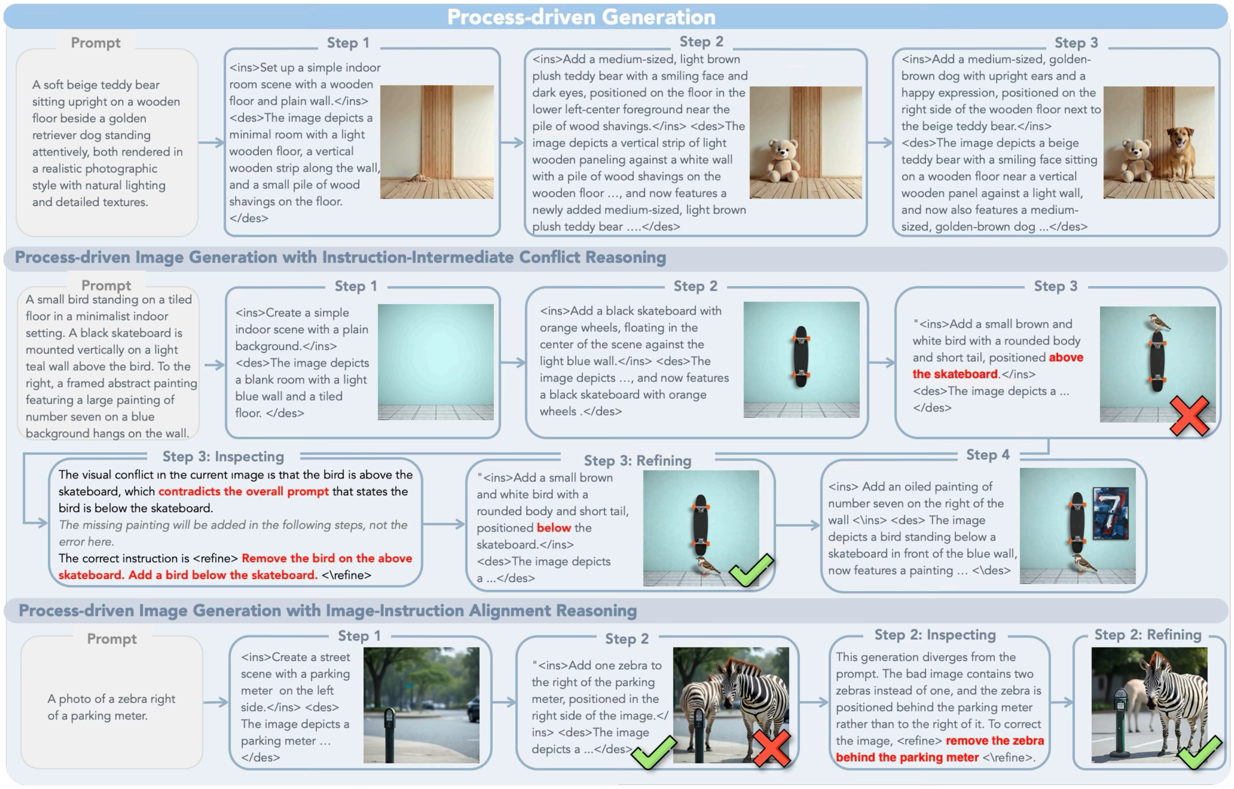

Process-driven Generation
Prompt
A soft beige teddy bear sitting upright on a wooden floor beside a golden retriever dog standing attentively, both rendered in a realistic photographic style with natural lighting and detailed textures.
Step 1
<ins>Set up a simple indoor room scene with a wooden floor and plain wall.</ins> <des>The image depicts a minimal room with a light wooden floor, a vertical wooden strip along the wall, and a small pile of wood shavings on the floor. </des>
Step 2
<ins>Add a medium-sized, light brown plush teddy bear with a smiling face and dark eyes, positioned on the floor in the lower left-center foreground near the pile of wood shavings.</ins> <des>The image depicts a vertical strip of light wooden paneling against a white wall with a pile of wood shavings on the wooden floor …, and now features a newly added medium-sized, light brown plush teddy bear ….</des>
Step 3
<ins>Add a medium-sized, golden-brown dog with upright ears and a happy expression, positioned on the right side of the wooden floor next to the beige teddy bear.</ins> <des>The image depicts a beige teddy bear with a smiling face sitting on a wooden floor near a vertical wooden panel against a light wall, and now also features a medium-sized, golden-brown dog ...</des>
Process-driven Image Generation with Instruction-Intermediate Conflict Reasoning
Prompt
A small bird standing on a tiled floor in a minimalist indoor setting. A black skateboard is mounted vertically on a light teal wall above the bird. To the right, a framed abstract painting featuring a large painting of number seven on a blue background hangs on the wall.
Step 1
<ins>Create a simple indoor scene with a plain background.</ins> <des>The image depicts a blank room with a light blue wall and a tiled floor. </des>
Step 2
<ins>Add a black skateboard with orange wheels, floating in the center of the scene against the light blue wall.</ins> <des>The image depicts …, and now features a black skateboard with orange wheels .</des>
Step 3
"<ins>Add a small brown and white bird with a rounded body and short tail, positioned above the skateboard.</ins> <des>The image depicts a ... </des>
Step 3: Inspecting
The visual conflict in the current image is that the bird is above the skateboard, which contradicts the overall prompt that states the bird is below the skateboard.
The missing painting will be added in the following steps, not the error here.
The correct instruction is <refine> Remove the bird on the above skateboard. Add a bird below the skateboard. <\refine>
Step 3: Refining
"<ins>Add a small brown and white bird with a rounded body and short tail, positioned below the skateboard.</ins> <des>The image depicts a ...</des>
Step 4
<ins> Add an oiled painting of number seven on the right of the wall <\ins> <des> The image depicts a bird standing below a skateboard in front of the blue wall, now features a painting … <\des>
Process-driven Image Generation with Image-Instruction Alignment Reasoning
Prompt
A photo of a zebra right of a parking meter.
Step 1
<ins>Create a street scene with a parking meter on the left side.</ins> <des> The image depicts a parking meter … </des>
Step 2
"<ins>Add one zebra to the right of the parking meter, positioned in the right side of the image.</ ins> <des>The image depicts a ...</des>
Step 2: Inspecting
This generation diverges from the prompt. The bad image contains two zebras instead of one, and the zebra is positioned behind the parking meter rather than to the right of it. To correct the image, <refine> remove the zebra behind the parking meter <\refine>.
Step 2: Refining

Figure 5 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 the instruction, where the instruction and overall prompt remain valid but the image requires refinement.

Ablation Study ③: Complementary Semantic- and visual-level constraints address distinct failure modes and jointly improve compositional reasoning accuracy.

Case	Counting	Colors	Position	Color Attri.
	0.61	0.84	0.66	0.58
w/ ins.	0.62	0.85	0.71	0.61
w/ img-ins.	0.73	0.86	0.69	0.61
w/ ins. + img-ins.	0.75	0.87	0.72	0.63

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

Table 7 analyzes two complementary mechanisms that intervene on intermediate reasoning states. Adding instruction-intermediate conflict supervision (w/ ins.) improves semantic and spatial consistency, evidenced by the gains on the Position category (+5%). This mechanism helps ensure that the evolving plan remains semantically aligned with the global intent, thereby reducing instruction drift and preserving correctness throughout the reasoning trajectory, which directly translates to better grounding in spatial and attribute-centric tasks. On the other hand, image-instruction alignment supervision (w/ img-ins.) sharpens visually grounded reasoning, producing substantial gains on Counting (+12%). Notably, combining both mechanisms achieves the highest performance across all tasks, revealing that semantic-level and visual-level intermediate checks target distinct and complementary failure modes. These results establish that enforcing correctness during intermediate steps—not merely at the final output—is essential for reliable multi-step generation, enabling the model to actively detect and rectify inconsistencies before they propagate.

4.5 Qualitative Evaluation

Figure 5. 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 distinct types of intermediate errors: (1) conflicts between the step-level instruction and the overall prompt, and (2) mismatches between the instruction and the partially generated image. These complementary mechanisms enable the model to revise wrong intermediate states and supply 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. As shown in Figure 6., our process-driven approach produces images with high visual fidelity, fine-grained details, and strong aesthetic appeal. The prompts are sampled from Gen-Eval and WISE benchmark.

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 own errors, and end-to-end training of BAGEL-7B to autoregressively emit interleaved text and image tokens. We lift the public BAGEL-7B from 0.77 to 0.82 (+5% absolute gain) on GenEval, while emergent abilities arise, including information-enriched reasoning and complexity adaptive planning. Looking forward, we will extend unified multimodal reasoning to videos and 3D space, and enable real-time human-in-the-loop control, unlocks truly controllable, truthful and interpretable image synthesis.

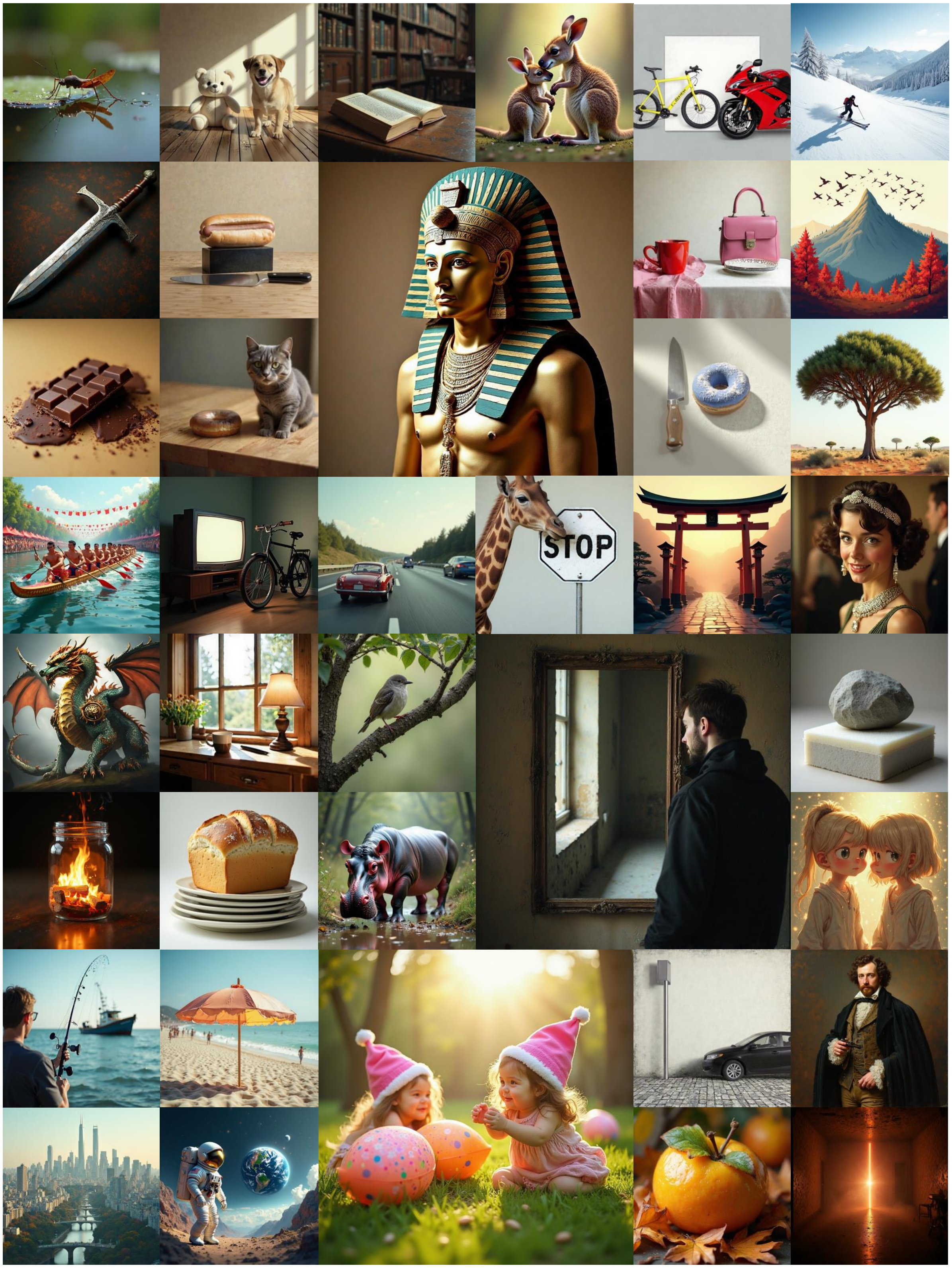


Figure 6 Visualization of generated image in our process-driven image generation. Our process-driven approach produces images with high visual fidelity, fine-grained details, and strong aesthetic appeal. The prompts are sampled from Gen-Eval and WISE benchmark.

References

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. https://api.semanticscholar.org/CorpusID:264403242.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\sigma : Weak-to-strongtrainingofdiffusiontransformerfor4ktext-to-imagegeneration$, 2024.

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, 2025. https://arxiv.org/abs/2501.17811.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022. https://arxiv.org/abs/2205.09712.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and Aixin Liu et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. https://arxiv.org/abs/2501.12948.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation, 2024. https://arxiv.org/abs/2309.11499.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. https://arxiv.org/abs/2403.03206.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing, 2025. https://arxiv.org/abs/2503.10639.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *EMNLP*, pages 1295–1309, November 2020.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2025. https://arxiv.org/abs/2404.14396.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. https://arxiv.org/abs/2310.11513.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step, 2025. https://arxiv.org/abs/2501.13926.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models, 2024. https://arxiv.org/abs/2406.09403.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. https://arxiv.org/abs/2503.06749.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot, 2025. https://arxiv.org/abs/2505.00703.

Ji Hyeok Jung, Eun Tae Kim, Seoyeon Kim, Joo Ho Lee, Bumsoo Kim, and Buru Chang. Is 'right' right? enhancing object orientation understanding in multimodal large language models through egocentric instruction tuning, 2025. https://arxiv.org/abs/2411.16761.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. https://arxiv.org/abs/2506.15742.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought, 2025a. https://arxiv.org/abs/2501.07542.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. https://arxiv.org/abs/2402.17245.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations, 2025b. https://arxiv.org/abs/2506.04633.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning, 2025. https://arxiv.org/abs/2503.19312.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2025. https://arxiv.org/abs/2402.08268.

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

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2025. https://arxiv.org/abs/2411.07975.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain of thought prompting for large multimodal models. In *CVPR*, 2024.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought, 2023. https://arxiv.org/abs/2305.15021.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation, 2025. https://arxiv.org/abs/2503.07265.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, and Alexander Wei et al. Openai o1 system card, 2024. https://arxiv.org/abs/2412.16720.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. https://arxiv.org/abs/2504.06256.

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, 2023. https://arxiv.org/abs/2307.01952.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2025. https://arxiv.org/abs/2508.05606.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation, 2025. https://arxiv.org/abs/2412.03069.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022a. https://arxiv.org/abs/2112.10752.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022b. https://arxiv.org/abs/2112.10752.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation, 2025. https://arxiv.org/abs/2412.15188.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. https://arxiv.org/abs/2405.09818.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning, 2024. https://arxiv.org/abs/2412.14164.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. https://arxiv.org/abs/1711.00937.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance, 2024a. https://arxiv.org/abs/2412.06673.

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

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting, 2024. https://arxiv.org/abs/2402.10200.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. https://arxiv.org/abs/2201.11903.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. https://arxiv.org/abs/2410.13848.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2024b. https://arxiv.org/abs/2309.05519.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025a. https://arxiv.org/abs/2409.04429.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms, 2025b. https://arxiv.org/abs/2502.07266.

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, 2025a. https://arxiv.org/abs/2408.12528.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025b. https://arxiv.org/abs/2506.15564.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl, 2025. https://arxiv.org/abs/2505.24875.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS2023*, 2023.

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, 2024a. https://arxiv.org/abs/2408.11039.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models, 2024b. https://arxiv.org/abs/2405.13872.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *ICCV*, 2025.