MetaPoint: Unlocking Precise Spatial Control in Agentic Visual Generation

Dewei Zhou[1,2*], Xinyu Huang[2*], Xun Wang[2*], Ji Xie[1,2], Yabo Zhang[2], Liang Li[2], Kunchang Li[2], Zongxin Yang[3], and Yi Yang[1†]

[1] Zhejiang University
[2] ByteDance Seed
[3] Harvard University
yangyics@zju.edu.cn
[*]Equal contribution. [†]Corresponding author. Xun Wang is the project leader.

Abstract. Generative visual models fundamentally struggle with precise spatial control. This arises from a core disconnect: models can process textual descriptions of space but cannot directly map numerical coordinates onto the 2D image canvas (as illustrated in Fig. 2). We introduce **MetaPoint**, a method that bridges this gap by representing a continuous 2D coordinate as a single, special token. Crucially, MetaPoint requires no new architectural components; it directly leverages the model's inherent positional encoding schemes to interpret these coordinates, treating our token as a virtual point on the canvas. This lightweight approach enables pixel-level control of an object's position with one token or its bounding box with two, all without requiring architectural changes or bespoke attention masking. The MetaPoint tokens are designed to be compositional, serving as spatial primitives. This allows a planner agent to decompose a high-level user request into a structured sequence of primitives for the generator. By providing a simple, precise, and scalable building block for spatial control, MetaPoint unlocks more powerful compositional generative agents and enables intuitive, interactive editing systems.

Keywords: Spatial control · image generation · diffusion models

1 Introduction

Unified Multimodal Models (UMMs) [7, 14, 40, 45], which unify generation and understanding across text and image modalities, have achieved notable success in precisely rendering photorealistic imagery from complex user prompts. Earlier advances in Large Language Models (LLMs) showed that purely text-based generation–understanding models can easily parse and reason over numerical coordinates in prompts [11, 44]. This naturally motivates the question of whether UMMs, with their integrated visual generation capabilities, can not only understand spatial specifications in text but also translate that understanding into precise visual layouts. In practice, however, a notable gap emerges: even for simple instructions, current UMMs often fail to place or locate objects accurately

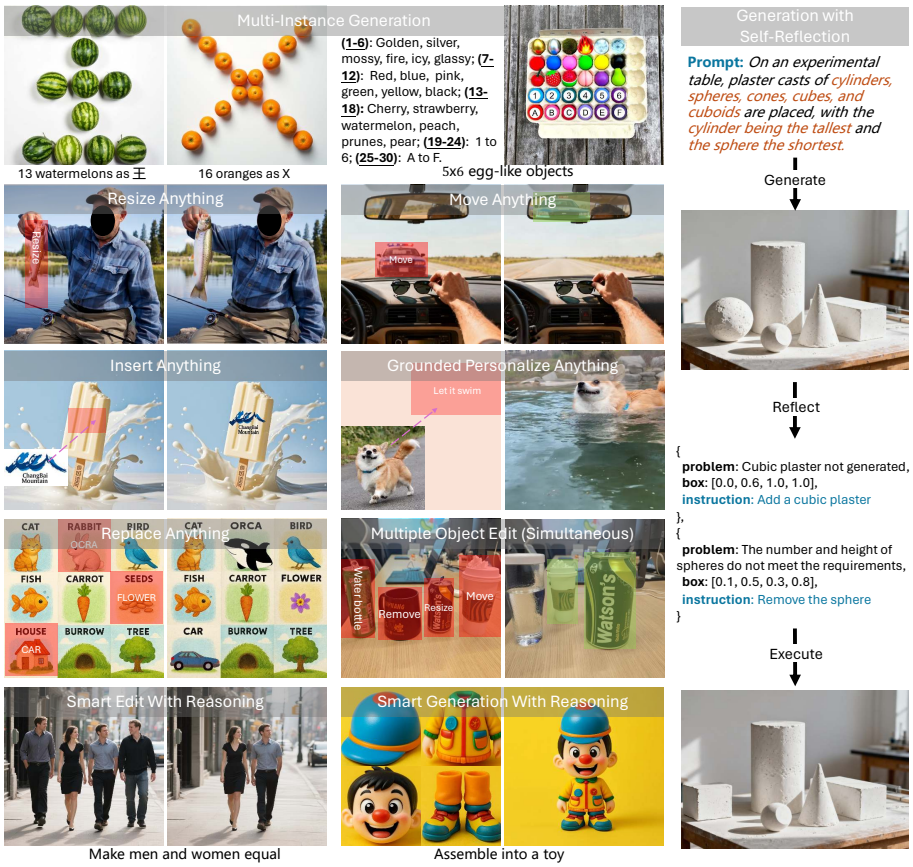

Fig. 1: MetaPoint demonstrates a wide array of capabilities, including complex **multi-instance generation**, versatile object editing (**move**, **insert**, **resize**, **replace**, and **multi-object editing**), and high-level **smart generation** with reflection from VLM-agent.

at the specified coordinates on the canvas (see Fig. 2). This gap between textual spatial specifications and the rendered visual outcome limits their reliability in applications that demand strict spatial accuracy.

Researchers have worked hard to improve position control, yet existing methods still have clear limitations (see Tab. 1). For instance, attention-masking approaches [61, 74] provide only coarse, patch-level control. Adapter-based methods [28, 50, 62, 77] are difficult to integrate into a single unified model. Position-vocabulary methods, such as ReCo [59], rely on a large discrete vocabulary of position tokens, which increases computational and memory cost and prevents pixel-accurate specification of continuous coordinates. In short, a persistent trade-off remains among accuracy, scalability, and architectural simplicity.

Fig. 2: Given a coordinate-based editing instruction, GPT-4o-Image [40] and Nano-Banana [14] both fail to precisely follow simple coordinates, exposing a key barrier to exact grounded generation. In contrast, our method (**MetaPoint**) accurately locates the specified region and performs the generation correctly.

*A **lightweight, scalable, model-agnostic** solution that offers precise control remains an open challenge.*

Most modern generative models [7, 14, 22, 24, 30, 40, 45], including UMMs, employ positional encoding schemes to represent locations in 2D visual space (Eqs. 1, 3). Motivated by this, we introduce MetaPoint: a special textual token (`<mp>`) equipped with spatial positional encoding. The token's word embedding conveys the intent to control position, while its positional embedding is engineered to encode the exact coordinates of a target point (see Fig. 3). This lightweight approach delivers pixel-level precision without architectural surgery, heavyweight vocabulary expansion, or coarse approximations.

By integrating this primitive into UMMs, we unlock a powerful compositional system for spatial control: a single MetaPoint specifies a point; a pair defines a precise bounding box; and short sequences encode complex poses, trajectories, or object layouts. This transforms the UMM from a system that merely responds to ambiguous text into a precise surgical tool. It can now follow instructions augmented with MetaPoint primitives to perform generation and editing with exact spatial awareness. Finally, to harness this compositional power and make it accessible for complex or casually phrased user requests, we introduce the MetaPoint-Agent in Fig. 3. This VLM-based planner translates high-level user intent into a sequence of clear, explicit MetaPoint-based instructions and acts as a crucial interpreter, converting ambiguous human language into precise, machine-executable commands that our enhanced UMM can execute.

Empirically, MetaPoint achieves strong and consistent gains on diverse benchmarks: it improves mIoU on COCO-MIG [75] from 59.23% to 77.29% (+**30.49%** relative) compared to the prior SOTA, boosts BAGEL's overall score on T2I-CoReBench [27] from 38.2 to 66.1 (+**73%**), and raises ImgEdit [60] overall from 3.42 to 3.94 (+**15.2%**). Notably, the advantage grows with task difficulty (e.g., higher object counts), suggesting that explicit 2D positional grounding is a key inductive bias for robust spatial reasoning.

Beyond benchmarks, MetaPoint shows two forms of scalability (Fig. 1): (i) reliable layout control for scenes with up to 30 objects while preserving visual fidelity; and (ii) simultaneous, coordinated editing of multiple objects of different types (*e.g.*, replace, resize, move, and remove) without altering non-edited regions. Looking forward, MetaPoint —as a stable, compositional primitive paired

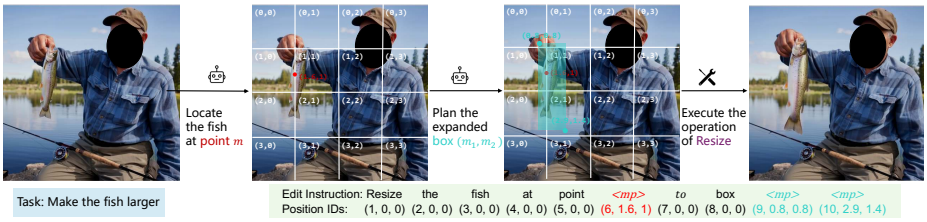

Fig. 3: MetaPoint and its agent. (1) **MetaPoint**, which uses special `<mp>` tokens linked to continuous 2D positional embeddings to localize targets (e.g., a fish) and to depict a guiding target box for the diffusion backbone; the `<mp>` positional embeddings are aligned with both the textual sequence position IDs and the 2D spatial coordinates of the corresponding pixels. (2) **MetaPoint-Agent**, a Vision-Language Model (VLM), which decomposes complex tasks into sub-tasks, plans a solution pathway, and generates executable natural commands to drive the MetaPoint model.

with a capable agent planner—forms a foundation for reliable, interactive generative systems that close the gap between numeric spatial intent and precise visual execution.

Our contributions are summarized as follows:

- We propose MetaPoint, a lightweight, model-agnostic, single-token interface that reuses the UMM's native positional encoding to achieve pixel-level spatial control—without architectural modifications, heavyweight vocabulary expansion, or coarse approximations.
- We show that MetaPoint tokens are naturally compositional: a single token specifies a point, a pair defines a bounding box, and sequences encode complex layouts, or editing operations. Paired with a VLM-based planner (MetaPoint-Agent), this enables an end-to-end system that translates high-level user intent into precise, spatially grounded generation and editing.
- Extensive experiments demonstrate that MetaPoint establishes new state-of-the-art results on COCO-MIG, T2I-CoReBench, and ImgEdit, with gains that grow with task complexity, and scales reliably to scenes with up to 30 objects and multi-object coordinated editing.

2 Related Work

2.1 Precise Spatial Control in Image Generation

Achieving precise spatial control in diffusion models has been a significant research focus. Early approaches relied on modifying the generation process with explicit guidance or masks. A prominent line of work involves **injecting spatial conditions via dedicated modules**. Methods like GLIGEN [23, 28, 50, 62, 75, 77, 78] introduce attention mechanisms or feature injection modules directly into the UNet or DiT backbone. While effective, these methods require **substantial architectural modifications**, making them heavyweight and difficult to integrate into evolving model architectures. ReCo [59] introduces 1,000

position-specific text tokens into the CLIP [41] text encoder. However, it brings heavyweight vocabulary expansion and significant training overhead. Moreover, its discrete token design is **fundamentally imprecise and unscalable**, making it impossible to specify continuous coordinates, and it cannot extend to high-resolution generation without an exponential increase in its already-large token set.

2.2 Agent-based Visual Generation

Recent advancements in generative models [2, 7, 8, 14, 18, 34–37, 42, 45, 65, 67–71, 73, 76, 79] have enabled impressive capabilities. However, they consistently fail on tasks requiring high compositional fidelity and precise spatial control [1, 12, 16, 17, 20, 26, 38, 39, 57]. To mitigate these reasoning and compositional limitations, agent-based frameworks have emerged. **Prompt rewriting methods** employ an LLM to revise the user's initial prompt into a more detailed and suitable version for the generative model [7, 19, 31, 48, 51], but this remains a text-only solution and is fundamentally incapable of fulfilling requests that require explicit spatial control. **Tool-calling methods** equip the agent with a "toolbox" of specialized and separated tools [10, 25, 52, 53, 56, 58, 64] that are difficult to synergize and collaborate, leading to error propagation, poor controllability, and a failure to achieve stronger emergent functions, such as inserting a user-specific object into a given image. Our proposed MetaPoint provides the *explicit spatial control* that text-only solutions lack, while its *native, unified design* avoids the synergy failures and error propagation inherent in fragmented toolboxes, providing a foundational component for a truly capable generative agent.

3 MetaPoint Method

Controlling object locations precisely and efficiently is central to generative AI. An ideal interface should (i) achieve pixel-level accuracy, (ii) use as few tokens as possible to avoid sequence-length and latency penalties, and (iii) plug into existing Unified Multimodal Models (UMMs) without architectural changes. Prior approaches do not meet all three goals simultaneously.

As shown in Fig. 3, we introduce MetaPoint, a single-token interface for pixel-level spatial control. The core idea is to reuse the UMM's native image positional encoding to directly represent a continuous 2D coordinate with one special token `<mp>`. This design is minimal, token-efficient, and architecture-agnostic. We first revisit positional encoding in UMMs (§3.1), then present the MetaPoint token (§3.2), and show how it enables spatial primitives (§3.3).

3.1 Positional Encoding in UMMs

UMMs reason over tokens embedded in a 3D index space: a sequence axis for order and two spatial axes (height, width) for layout. Two families of positional encoding are common: a hybrid scheme using 1D RoPE [43] for sequence and

Table 1: Comparison of position-control methods. MetaPoint uniquely achieves pixel-level precision, single-token efficiency, and native UMM compatibility. We evaluate precision (Prec.), vocab token additions, and UMM compatibility.

Method	Prec.	Add Tokens	UMM
Text Prompt [14, 40, 45]	*roughly*	0	✔
Attn Mask [61, 74]	*patch*	0	✗
Adapter [62, 77]	*pixel*	0	✗
Position Vocabulary [59]	≈ *pixel*	N	✗
MetaPoint	*pixel*	1	✔

2D Sinusoidal Positional Embedding [46] for space, and a unified 3D RoPE that applies rotary embeddings across all axes. The main difference is whether sequence and spatial positions are handled by separate or shared mechanisms.

2D Sinusoidal PE. Let (u, v) denote the spatial coordinates of an image token. The 2D sinusoidal positional embedding maps (u, v) to a d-dimensional vector by splitting the dimensions into two halves: the first $d/2$ encode the row u, and the second $d/2$ encode the column v,

$$\mathrm{PE}_u(i) = \sin\left(\frac{u}{K^{4i/d}}\right), \quad \mathrm{PE}_u(i + d/4) = \cos\left(\frac{u}{K^{4i/d}}\right), \tag{1}$$

where $i \in \{0, 1, \ldots, d/4 - 1\}$ and K is a base (e.g., 10000). The second half encodes v in the same way, and the final embedding is the concatenation:

$$\mathrm{PE} = [\mathrm{PE}_u, \mathrm{PE}_v]. \tag{2}$$

The visual token embedding is augmented by adding PE.

3D RoPE. In self-attention [47], Rotary Position Embedding (RoPE) [43] applies position-dependent rotations. For UMMs, the embedding dimension d is partitioned into three blocks for the sequence position, height, and width ($d = d_1 + d_2 + d_3$). Independent rotations are applied to each block with respect to the corresponding index m:

$$\mathcal{R}(m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix}. \tag{3}$$

For a visual token at (u, v), rotations $\mathcal{R}(u)$ and $\mathcal{R}(v)$ are applied to the spatial chunks using a list of angles θ. See BAGEL [7] and Mogao [30] for details.

3.2 MetaPoint

UMMs already encode spatial relations via their positional systems. We ask: how can we directly leverage this native capability to represent a specific 2D location with maximal precision and minimal overhead?

We define a single special token, `<mp>`, that serves as a pointer to a continuous 2D coordinate (u, v). When `<mp>` is used, the UMM applies its existing spatial

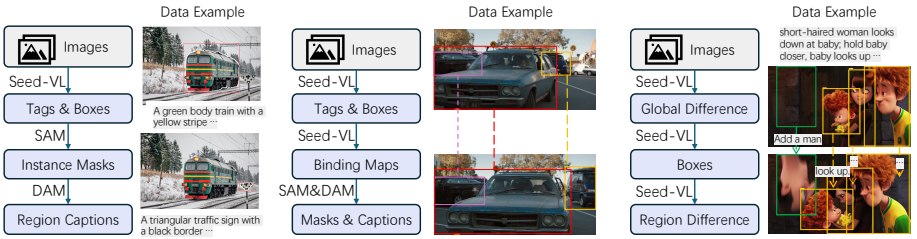

Fig. 4: Point-Anchored Data Pipeline. (a) Caption Layout: Single image; tags/-captions/boxes/masks; bind text captions to MetaPoints. **(b) Exemplar Insert:** Paired frames; boxes and binding maps for the same object between two frames; bind visual exemplar to MetaPoint. **(c) Instruction Edit:** Paired frames; global diffs → regional edit instruction; bind MetaPoints to multi-region edits.

positional encoding (e.g., 2D Sinusoidal PE or 3D RoPE) to (u, v) as if it were an image token[4], as illustrated in Fig. 3. Concretely, the embedding of MetaPoint at (u, v) is

$$\mathrm{X}_{u,v} = \mathrm{X}_{\texttt{<mp>}} + [\mathrm{PE}_u, \mathrm{PE}_v], \tag{4}$$

where $\mathrm{X}_{\texttt{<mp>}}$ is the learned vocabulary embedding for the special token.

Crucially, we treat (u, v) as continuous coordinates. Although positional encodings are often applied to discrete patch indices, the formulas in (1), (2), and (3) are differentiable and accept floating-point inputs. Supplying continuous (u, v) breaks free from the patch grid and enables pixel-level control independent of the generation resolution.

Intuitively, MetaPoint models a virtual "click" on the canvas: the click extracts the exact location and compresses it into one token. The design is notable for its simplicity and effectiveness: (i) one token per point, (ii) no architectural changes, and (iii) direct reuse of the model's spatial reasoning. MetaPoint attains both precision and efficiency without compromise.

3.3 Versatile Spatial Control via MetaPoint

MetaPoint enables versatile spatial primitives (Fig. 1), and this section shows how MetaPoint can be composed to achieve spatial control in different tasks, from generation to editing, and how they integrate with a planner (e.g., a VLM) for instruction-driven workflows and self-reflective autonomous correction:

– **Generation: layout and count.** A single token `<mp>` can specify where an object should be generated. Two tokens $[\texttt{mp}_{\mathrm{tl}}, \texttt{mp}_{\mathrm{br}}]$ define a bounding box, allowing both position and size control. A sequence of N tokens $[\texttt{mp}_1, \ldots, \texttt{mp}_N]$ specifies the layout of N objects.

[4] BAGEL uses 2D Sinusoidal PE, and all results in the paper use MetaPoint with the same PE. We also implemented MetaPoint on an in-house UMM with a 3D RoPE variant.

Table 2: Statistics of the training datasets. We mix some in-house datasets (first block) with our newly constructed datasets (second block) and report their sampling ratios. Acronyms (PACL, PAEI, PAIE) are defined in Sec. 4.

Dataset	# Samples	Ratio	Supporting Task
T2I Data	1M	3.8%	Maintain t2i generation
OCR Data	50K	0.2%	Text edit
PACL	3M	31%	Caption-driven layout
PAEI	3M	35%	Exemplar-based insertion
PAIE	2M	30%	Instruction-driven edit

- **Object editing.** Edits can be conditioned on a *source* box alone or on a *source–target* pair. In the first case, a token sequence such as $[\texttt{mp}_{\text{src1}}, \texttt{mp}_{\text{src2}}]$ specifies the box of the object to be modified, while the desired change (e.g., recolor, replace) is described in natural language. In the second case, both a source and a target box are provided, e.g., $[\texttt{mp}_{\text{src1}}, \texttt{mp}_{\text{src2}}] \rightarrow [\texttt{mp}_{\text{tgt1}}, \texttt{mp}_{\text{tgt2}}]$, instructing the model to transform the source object to fit the target region—moving, resizing, or modifying pose. This flexible mechanism supports a wide range of spatial edits, including insertion, replacement, moving, and resizing.
- **Integration with a VLM Planner.** A VLM [15] bridges user intent and low-level MetaPoint primitives. As shown in Fig. 3, the VLM (1) **perceives** the image and localizes the fish as $\texttt{mp}_{\text{src}}$, (2) **reasons** about the intent to plan a target box $[\texttt{mp}_{\text{t1}}, \texttt{mp}_{\text{t2}}]$, and (3) **generates** the natural language instruction based on MetaPoint primitives.
- **Self-reflection and Autonomic Correction.** The precision of MetaPoint-based editing enables a VLM Planner not only to generate and edit content, but also to *self-reflect* on its own outputs. After generating an image, the VLM can re-analyze the result, localize specific spatial errors (e.g., missing objects, incorrect size or count), and solve each problem with exact bounding boxes via MetaPoint tokens. It can then issue corrective edit instructions, such as "add a cubic plaster" or "remove the sphere" with precise coordinates (Fig. 1). This closed-loop process—*generate* $\rightarrow$ *reflect* $\rightarrow$ *execute*—enables autonomous improvement without human intervention.

4 Data Pipeline

Video provides rich supervision via temporal consistency and variation. Harnessing this principle, our data engine automatically constructs three distinct datasets to enable precise spatial control, *e.g.*, layout, insertion, and editing. For brevity, we denote these datasets as **PACL** (Point-Anchored Caption Layout), **PAEI** (Point-Anchored Exemplar Insert), and **PAIE** (Point-Anchored Instruction Edit) (see Fig. 4 for an overview).

PACL. We generate layout data with dense supervision (Fig. 4*a*): Seed-VL [15] yields precise tags and tight boxes, SAM [21] provides pixel masks, and DAM [29]

Table 3: Quantitative comparison on **COCO-MIG** for multi-instance generation tasks. **Instance Success Rate (%)** ↑ and **mIoU** ↑ metrics across different object count levels (L_2 to L_6).

Method	Instance Success Rate(%) ↑						mIoU ↑					
	Avg	L_2	L_3	L_4	L_5	L_6	Avg	L_2	L_3	L_4	L_5	L_6
LAMIC [5]	13.56	28.12	19.17	13.75	9.00	9.58	21.17	31.67	25.79	20.68	18.08	18.25
GrounDiT [23]	22.91	36.56	31.25	22.97	17.75	18.44	29.72	37.41	35.30	30.13	26.50	26.79
GLIGEN [28]	29.56	41.88	31.67	27.19	27.38	27.81	27.44	37.35	29.17	25.31	26.42	25.56
MS-Diffusion [49]	28.22	37.81	33.12	28.12	25.75	24.69	34.69	41.15	36.38	34.57	32.36	33.70
CreatiLayout [63]	54.69	67.19	63.33	56.09	50.25	48.96	48.96	56.32	55.38	49.42	46.22	45.28
InstanceDiffusion [50]	60.28	71.25	61.67	59.38	57.00	59.27	54.79	65.76	57.21	53.33	51.43	53.72
ReCo [59]	56.90	65.50	56.10	56.30	52.40	58.30	47.60	55.70	46.70	47.20	43.30	48.80
EliGen [61]	64.12	69.69	72.50	66.56	61.62	58.54	59.23	64.61	66.10	61.59	56.74	54.50
MIGC [75]	66.44	74.06	67.29	67.03	63.25	65.73	56.96	63.84	57.60	56.95	54.01	56.82
BAGEL+MetaPoint	**84.72**	**84.52**	**84.31**	**86.66**	**83.29**	**84.85**	**77.29**	**76.72**	**76.60**	**79.32**	**76.22**	**77.34**
vs. SOTA	+18.28	+10.46	+11.81	+19.63	+20.04	+19.12	+18.06	+10.96	+10.50	+17.73	+19.48	+20.52

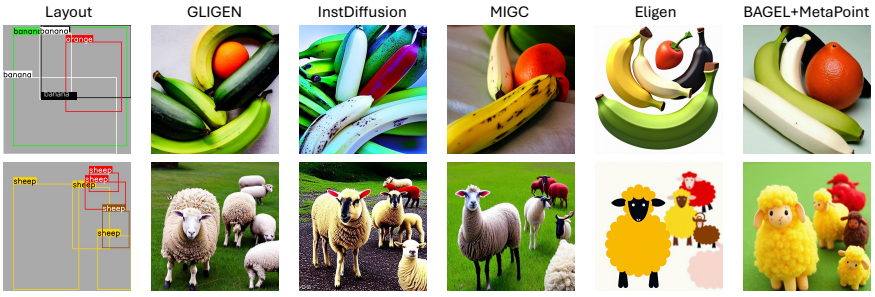

Fig. 5: Qualitative Results on COCO-MIG benchmark. Each instance is assigned a location and color, shown by its bounding box.

produces region captions. Each image is annotated with tags, boxes, masks, and captions anchored to MetaPoint.

PAEI. For visual exemplar insertion (Fig. 4*b*), we sample frame pairs from videos, using Seed-VL to detect objects and establish correspondences. For training, one frame serves as the ground truth (GT) frame. Objects from the other frame, specified by their bounding box in the GT frame, become the visual exemplar (grounded image condition). The model learns to insert this visual exemplar at a location specified by the MetaPoint.

PAIE. For editing (Fig. 4*c*), we pair frames from videos, detect global and region-level changes with Seed-VL, and auto-generate concise edit instructions (*e.g.*, add, move, resize, remove) bound to MetaPoint. This trains precise, localized edits while preserving the background.

5 Experiments

5.1 Experimental Setup

Implementation Details. We train our model, MetaPoint, on the datasets detailed in Tab. 2 for 10K steps, starting from BAGEL [7]. The training is conducted on 256 H20 GPUs and takes approximately two days to complete.

Table 4: Main results on T2I-CoReBench assessing both *composition* and *reasoning* capabilities.

Model	Composition					Reasoning									Overall
	MI	MA	MR	TR	Mean	LR	BR	HR	PR	GR	AR	CR	RR	Mean	
Closed-Source Models															
GPT-4o-Image [40]	84.1	75.9	72.7	86.4	79.8	59.0	54.8	65.6	87.3	76.5	82.0	70.9	56.1	69.0	72.6
Seedream 4.0 [45]	91.5	84.5	75.0	93.6	86.1	76.3	54.1	60.7	85.8	85.9	77.1	71.6	47.9	69.9	75.3
Nano Banana [14]	85.7	77.9	72.6	86.3	80.6	64.5	64.9	67.1	85.2	84.1	83.1	71.3	68.7	73.6	75.9
Imagen 4 Ultra [13]	90.0	80.0	73.2	86.2	82.4	63.6	62.4	66.1	88.5	82.8	83.0	76.3	60.7	72.9	76.1
Open-Source Models															
Janus-Pro-7B [4]	54.4	59.3	40.9	7.5	40.5	19.8	20.9	34.6	22.4	11.5	30.4	8.7	9.8	19.8	26.7
SD-3.5-Large [9]	57.5	60.0	32.9	15.6	41.5	22.5	22.4	34.2	52.5	35.5	53.0	42.3	25.2	35.9	37.8
BAGEL [7]	64.9	65.2	45.8	9.7	46.4	23.4	21.9	33.0	51.6	31.2	50.4	32.4	29.3	34.1	38.2
OmniGen2-7B [55]	67.9	64.1	48.3	19.2	49.9	24.7	23.2	43.3	63.1	46.1	54.2	36.5	24.1	39.4	42.9
HiDream-I1 [3]	62.5	62.0	42.9	33.9	50.3	34.2	24.5	40.9	53.2	34.2	50.3	46.1	31.7	39.4	43.0
FLUX.1-Krea-dev [22]	70.7	71.1	53.2	28.9	56.0	30.3	26.1	44.5	70.6	50.5	57.5	46.3	28.7	44.3	48.2
Qwen-Image [54]	**81.4**	**79.6**	**65.6**	**85.5**	**78.0**	41.1	32.2	48.2	75.1	56.5	53.3	**61.9**	26.4	49.3	58.9
BAGEL + MetaPoint	79.4	72.7	54.7	48.4	66.3	**73.3**	**53.4**	**62.2**	**80.5**	**79.1**	**67.0**	56.6	**55.9**	**66.0**	**66.1**
vs. BAGEL	+14.5	+7.5	+8.9	+38.7	+19.9	+49.9	+31.5	+29.2	+28.9	+47.9	+16.6	+24.2	+26.6	+31.9	+27.9

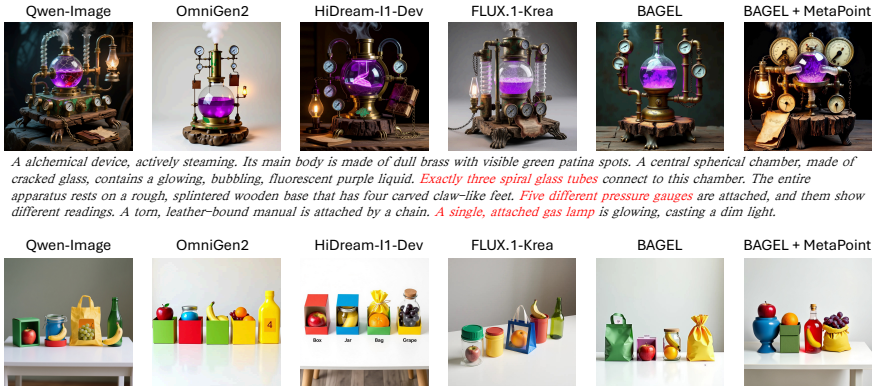

Fig. 6: Qualitative Results on T2I-CoReBench.

For a fair comparison during inference, our method strictly reuses the original configuration of BAGEL, including its resolution, sampler, steps, and CFG scale.

Evaluation. We evaluate our method on three benchmarks that cover both generative and editing tasks.

- **COCO-MIG** [75]: Layout-to-image benchmark evaluating a model's ability to control object placement and attribute bindings in multi-instance scenes.
- **T2I-CoReBench** [27]: Text-to-image benchmark for assessing a model's composition and reasoning capabilities.
- **ImgEdit** [60]: Image editing benchmark that measures performance across nine instruction-driven categories.

Table 5: Comparison results on ImgEdit.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
UltraEdit [72]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
ICEdit [66]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [33]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
UniWorld-V1 [32]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
BAGEL [7]	3.81	3.59	1.58	3.85	3.16	3.39	4.51	2.67	4.25	3.42
OmniGen2 [55]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
FLUX-Kontext [22]	3.83	3.65	2.27	4.45	3.17	3.98	4.55	3.35	4.29	3.71
GPT-4o-Image	**4.61**	**4.33**	2.90	4.35	3.66	**4.57**	**4.93**	**3.96**	**4.89**	4.20
Qwen-Image	4.38	4.16	**3.43**	**4.66**	4.14	4.38	4.81	3.82	4.69	**4.27**
BAGEL + MetaPoint	4.26	4.14	2.24	4.32	**4.20**	3.82	4.82	3.07	4.22	3.94
vs. BAGEL	+0.45	+0.55	+0.66	+0.47	+1.04	+0.43	+0.31	+0.40	-0.03	+0.52

For all benchmarks, we adapt a VLM to translate all tasks into a new instruction with `<mp>`[5] and then follow the original evaluation protocols. We report Instance Success Rate (%) ↑ (measuring attribute accuracy) and mIoU ↑ (measuring layout accuracy) for COCO-MIG, use the official Gemini 2.5 Flash [6] for scoring on T2I-CoReBench, and employ GPT-4.1 to score instruction alignment, visual fidelity, and realism (on a 1–5 scale) for ImgEdit.

5.2 Quantitative and Qualitative Results

COCO-MIG. As shown in Tab. 3, our method substantially outperforms the prior SOTA on COCO-MIG, raising the average Instance Success Rate from 66.44% to 84.72% (+**27.51%** relative) and mIoU from 59.23% to 77.29% (+**30.49%** relative). The advantage grows with task difficulty: as object count increases (L2→L6), Instance Success Rate gains expand from +10.46 (L2) to over +19 (L4–L6), and mIoU gains rise from +10.96 (L2) to +20.52 (L6). Performance is stable across levels (L2–L6), with Instance Success Rate at ≈84–87% and mIoU at ≈76–79%, indicating the benchmark's complexity is well within our model's capabilities. Qualitative results (Fig. 5) show more natural, coherent scenes that satisfy all constraints. While COCO-MIG is limited to 6 objects, MetaPoint scales well beyond the benchmark, controlling up to 30 objects with high fidelity (Fig. 1), demonstrating a capability that exceeds current methods.

T2I-CoReBench. Our method boosts the BAGEL baseline's overall score from 38.2 to 66.1, a **73%** relative improvement that establishes a new state-of-the-art for open-source models (Table 4). This improvement is driven by substantial gains in both Composition (mean score +19.9) and particularly in Reasoning (mean score +31.9). Notably, our method delivers its largest gains on sub-tasks where the baseline was particularly weak, such as Logical Reasoning (LR, +49.9), Geometric Reasoning (GR, +47.9), and Text Rendering (TR, +38.7). As shown in Fig. 6, these quantitative leaps are corroborated by qualitative results, where our model successfully handles nuanced prompts that challenge the baseline.

ImgEdit. As shown in Tab. 5, MetaPoint significantly enhances BAGEL, boosting the Overall score: **3.42** → **3.94** (+**15.2%** relative). While Qwen-Image achieves a higher Overall score (4.27), our method excels in specific tasks like

[5] The utilized system prompts are given in supplementary materials.

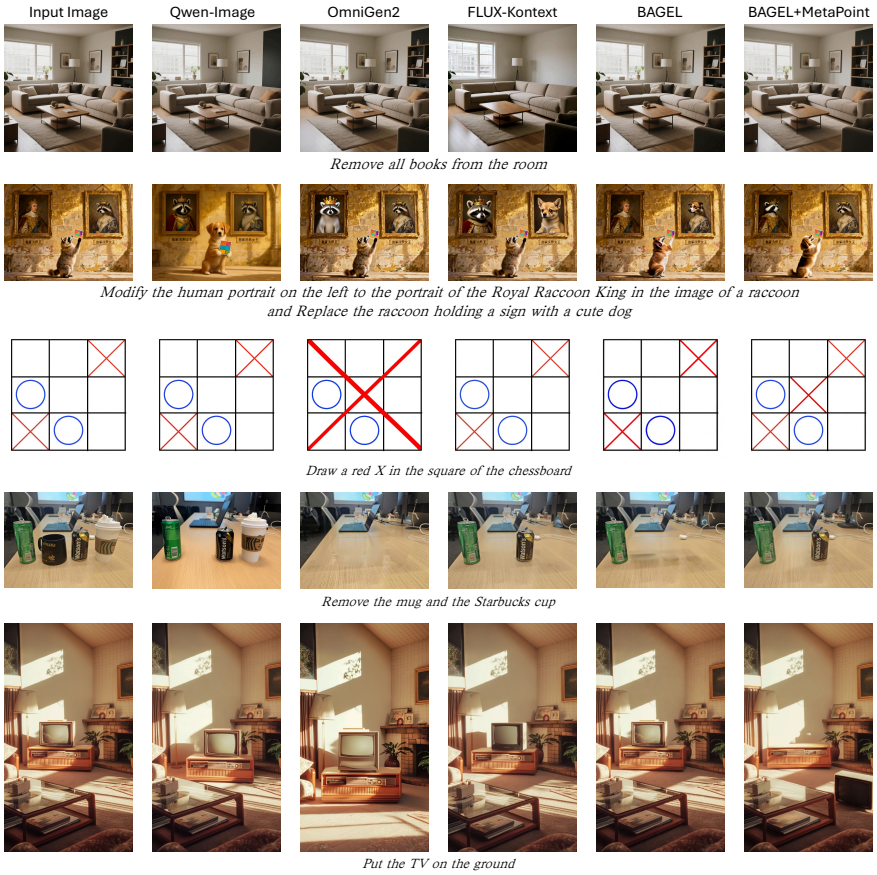

Fig. 7: Qualitative comparison on complex editing tasks.

Remove (3.16 → 4.20), where it achieves the best score. This performance is achieved using BAGEL with minimal edit-specific training, guided by a VLM agent that translates user requests into structured MetaPoint-based operations (*e.g.*, move, resize). Beyond benchmarks, as shown in Fig. 7, MetaPoint performs precise localized modifications while preserving the background, unlike text-only baselines that often cause grounding errors and unintended global changes.

5.3 Ablation of Positional Encoding

We ablate positional encoding while holding data, compute, and sampling constant. **BAGEL + Text** supplies natural-language coordinates (normalized to [0, 1000]) to the UMM and relies on the model to infer geometry from tokens; **BAGEL + MetaPoint** instead injects spatially aligned 2D positional embeddings directly into the DiT branch.

As reported in Tab. 6, replacing text coordinates with MetaPoint yields a large improvement on COCO-MIG: average Instance Success Rate rises from

Table 6: Ablation of MetaPoint. (Text *vs*. MetaPoint)

Method	Instance Success Rate(%) ↑						mIoU ↑					
	Avg	L_2	L_3	L_4	L_5	L_6	Avg	L_2	L_3	L_4	L_5	L_6
BAGEL+Text	61.84	50.00	61.44	60.30	62.14	66.77	52.48	47.78	52.69	51.86	52.31	54.48
BAGEL+MetaPoint	**84.72**	**84.52**	**84.31**	**86.66**	**83.29**	**84.85**	**77.29**	**76.72**	**76.60**	**79.32**	**76.22**	**77.34**

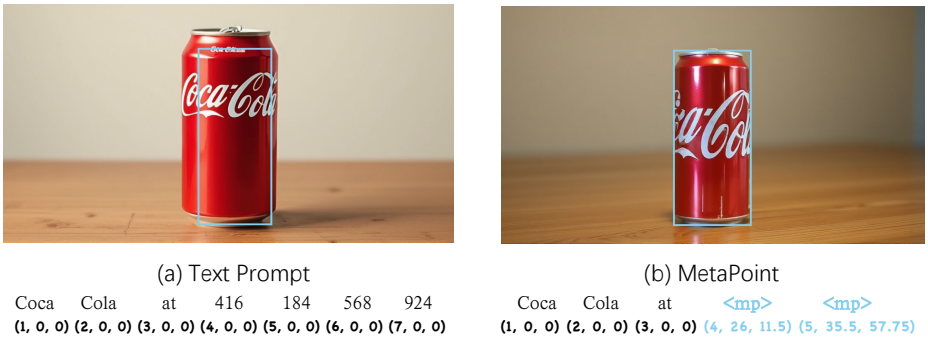

Fig. 8: Text *vs*. MetaPoint. Textual coordinates only yield coarse localization, while MetaPoint achieves precise, pixel-level placement using image-aligned 2D positional embeddings. In (b), the MetaPoint coordinates are BAGEL VAE token-space positions obtained by converting the original pixel box $(416, 184, 568, 924)$ with a /16 downsampling factor. The groundtruth box is highlighted in blue.

61.84% to 84.72% (+22.88) and average mIoU from 52.48% to 77.29% (+24.81). Notably, the text-based variant exhibits high variance across complexity levels (ISR ranges from 50.00 at L_2 to 66.77 at L_6), whereas MetaPoint remains consistently strong (≈83–87% ISR across all levels), indicating that explicit 2D positional encoding generalizes robustly regardless of scene complexity.

As shown in Fig. 8, the text-only variant preserves semantics but exhibits spatial drift, whereas MetaPoint anchors content at precise locations, achieving pixel-level control. The intuition is straightforward: converting coordinate meta-information into native 2D positional codes grants the model direct **visual** access to space, rather than requiring it to learn geometry through linguistic indirection.

We hypothesize that purely language-based learning could, in principle, approach similar precision, but would likely require substantially more compute, larger models, and much more data, and may still fall short of the generalization afforded by explicit 2D positional encoding. Task-aligned inductive bias, *i.e.*, choosing representations that match the structure of the problem, offers a practical path beyond scaling alone.

6 Conclusion and Limitations

MetaPoint enables precise spatial control in generative models through coordinate tokens, but three key limitations remain: (a) *Underutilized MetaPoint.* While MetaPoint handles basic operations (move, resize), richer controls like rotation remain unexplored through token extensions. (b) *Incomplete Control*

Tools. Position alone is insufficient; important attributes such as depth, pose, color, and texture still lack control. These capabilities form the foundation for truly controllable generation; without them, comprehensive and reliable control remains unattainable. (c) *Isolated Agent System.* The current agent's planning is based on manual system prompting, which prevents fluid tool integration. Future agents should dynamically assemble MetaPoint tokens with other tools to handle users' complex tasks. Solving these limitations will transform visual generation from "mysterious spell-casting" to "precise programming."

References

1. An, R., Yang, S., Guo, Z., Dai, W., Shen, Z., Li, H., Zhang, R., Wei, X., Li, G., Wu, W., et al.: Genius: Generative fluid intelligence evaluation suite. arXiv preprint arXiv:2602.11144 (2026)
2. Bader, J., Pach, M., Bravo, M.A., Belongie, S., Akata, Z.: Stitch: Training-free position control in multimodal diffusion transformers. arXiv preprint arXiv:2509.26644 (2025)
3. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
4. 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. arXiv preprint arXiv:2501.17811 (2025)
5. Chen, Y., Ma, Z., Wang, J., Kang, K., Yao, S., Zhang, W.: Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. arXiv preprint arXiv:2508.00477 (2025)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
8. Ding, Y., Zhuang, S., Li, K., Yue, Z., Qiao, Y., Wang, Y.: Muses: 3d-controllable image generation via multi-modal agent collaboration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2753–2761 (2025)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
10. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv:2503.10639 (2025)
11. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems **36**, 18225–18250 (2023)
12. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: NeurIPS (2023)

13. Google: Imagen 4. https://labs.google/fx/tools/image-fx (2025)
14. Google: Nano banana. https://gemini.google/overview/image-generation/ (2025)
15. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
16. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment (2024)
17. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
18. Huang, R., Cai, K., Han, J., Liang, X., Pei, R., Lu, G., Xu, S., Zhang, W., Xu, H.: Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In: European Conference on Computer Vision. pp. 144–160. Springer (2024)
19. 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. arXiv preprint arXiv:2505.00703 (2025)
20. Jin, W., Niu, Y., Liao, J., Duan, C., Li, A., Gao, S., Liu, X.: Srum: Fine-grained self-rewarding for unified multimodal models. arXiv preprint arXiv:2510.12784 (2025)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
22. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
23. Lee, Y., Yoon, T., Sung, M.: Groundit: Grounding diffusion transformers via noisy patch transplantation. Advances in Neural Information Processing Systems **37**, 58610–58636 (2024)
24. Li, H., Peng, X., Wang, Y., Peng, Z., Chen, X., Weng, R., Wang, J., Cai, X., Dai, W., Xiong, H.: Onecat: Decoder-only auto-regressive model for unified understanding and generation. arXiv preprint arXiv:2509.03498 (2025)
25. Li, H., Qing, C., Zhang, H., Jiang, D., Zou, Y., Peng, H., Li, D., Dai, Y., Lin, Z., Tian, J., et al.: Coco: Code as cot for text-to-image preview and rare concept generation. arXiv preprint arXiv:2603.08652 (2026)
26. Li, H., Li, Y., Lin, B., Niu, Y., Yang, Y., Huang, X., Cai, J., Jiang, X., Hu, Y., Chen, L.: Gir-bench: Versatile benchmark for generating images with reasoning. arXiv preprint arXiv:2510.11026 (2025)
27. Li, O., Wang, Y., Hu, X., Huang, H., Chen, R., Ou, J., Tao, X., Wan, P., Feng, F.: Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? arXiv preprint arXiv:2509.03516 (2025)
28. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
29. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video captioning. arXiv preprint arXiv:2504.16072 (2025)
30. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)

31. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17214–17223 (2025)
32. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
33. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
34. Lu, S., Lian, Z., Zhou, Z., Zhang, S., Zhao, C., Kong, A.W.K.: Does flux already know how to perform physically plausible image composition? arXiv preprint arXiv:2509.21278 (2025)
35. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: ICCV (2023)
36. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. CVPR (2024)
37. Lu, S., Zhou, Z., Lu, J., Zhu, Y., Kong, A.W.K.: Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint arXiv:2410.18775 (2024)
38. Niu, Y., Jin, W., Liao, J., Feng, C., Jin, P., Lin, B., Li, Z., Zhu, B., Yu, W., Yuan, L.: Does understanding inform generation in unified multimodal models? from analysis to path forward. arXiv preprint arXiv:2511.20561 (2025)
39. Niu, Y., Ning, M., Zheng, M., Lin, B., Jin, P., Liao, J., Ning, K., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
40. OpenAI: Gpt-4o. https://openai.com/index/introducing-4o-image-generation/ (2025)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
43. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
44. Team, O.: Omost github page (2024)
45. Team, S., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Wang, L., Xing, X., Cheng, Y., Zhao, Z., Li, D., Hang, T., Tao, J., Wang, Q., Li, R., Chen, C., et al.: Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. arXiv preprint arXiv:2509.04545 (2025)

49. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=PJqP0wyQek
50. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6232–6242 (2024)
51. Wang, Y., Liu, M., He, W., Zhang, L., Huang, Z., Zhang, G., Shu, F., Tao, Z., She, D., Yu, Z., et al.: Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. arXiv:2503.01298 (2025)
52. Wang, Z., Li, A., Li, Z., Liu, X.: Genartist: Multimodal llm as an agent for unified image generation and editing. Advances in Neural Information Processing Systems **37**, 128374–128395 (2024)
53. Wang, Z., Xie, E., Li, A., Wang, Z., Liu, X., Li, Z.: Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation. arXiv preprint arXiv:2401.15688 (2024)
54. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
55. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
56. Wu, T.H., Lian, L., Gonzalez, J.E., Li, B., Darrell, T.: Self-correcting llm-controlled diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6327–6336 (2024)
57. Xie, J., Darrell, T., Zettlemoyer, L., Wang, X.: Reconstruction alignment improves unified multimodal models. arXiv preprint arXiv:2509.07295 (2025)
58. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: International Conference on Machine Learning (2024)
59. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023)
60. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
61. Zhang, H., Duan, Z., Wang, X., Chen, Y., Zhang, Y.: Eligen: Entity-level controlled image generation with regional attention. arXiv preprint arXiv:2501.01097 (2025)
62. Zhang, H., Hong, D., Gao, T., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. arXiv preprint arXiv:2412.03859 (2024)
63. Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18487–18497 (2025)
64. Zhang, X., Yang, L., Li, G., Cai, Y., Xie, J., Tang, Y., Yang, Y., Wang, M., Cui, B.: Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. arXiv preprint arXiv:2410.07171 (2024)

65. Zhang, Y., Li, J., Tai, Y.W.: Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration. arXiv preprint arXiv:2504.00010 (2025)
66. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
67. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8281–8291 (2024)
68. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: Luve: Latent-cascaded ultra-high-resolution video generation with dual frequency experts. arXiv preprint arXiv:2602.11564 (2026)
69. Zhao, C., Chen, Z., Xu, Y., Gu, E., Li, J., Yi, Z., Wang, Q., Yang, J., Tai, Y.: From zero to detail: Deconstructing ultra-high-definition image restoration from progressive spectral perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17935–17946 (2025)
70. Zhao, C., Ci, E., Xu, Y., Fan, T., Guan, S., Ge, Y., Yang, J., Tai, Y.: Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. Advances in Neural Information Processing Systems **38**, 3373–3393 (2026)
71. Zhao, C., Xu, Y., Chen, Z., Gu, E., Zhang, K., Liu, X., Yang, J., Tai, Y.: From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
72. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)
73. Zhou, D., Li, M., Yang, Z., Lu, Y., Xu, Y., Wang, Z., Huang, Z., Yang, Y.: Bidedpo: Conditional image generation with simultaneous text and condition alignment. arXiv preprint arXiv:2511.19268 (2025)
74. Zhou, D., Li, M., Yang, Z., Yang, Y.: Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In: ICCV (2025)
75. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: CVPR (2024)
76. Zhou, D., Li, Y., Yang, Z., Yang, Y.: Refineanything: Multimodal region-specific refinement for perfect local details. arXiv preprint arXiv:2604.06870 (2026)
77. Zhou, D., Xie, J., Yang, Z., Yang, Y.: 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. arXiv preprint arXiv:2410.12669 (2024)
78. Zhou, D., Xie, J., Yang, Z., Yang, Y.: 3dis-flux: simple and efficient multi-instance generation with dit rendering. arXiv preprint arXiv:2501.05131 (2025)
79. Zhou, Z., Lu, S., Leng, S., Zhang, S., Lian, Z., Yu, X., Kong, A.W.K.: Dragflow: Unleashing dit priors with region based supervision for drag editing. arXiv preprint arXiv:2510.02253 (2025)