ConsistCompose: Unified Multimodal Layout Control for Image Composition

Supplementary Material

A. Training Recipe of ConsistCompose

We adopt a two-stage training recipe for ConsistCompose that injects layout awareness while preserving general multimodal understanding and generation capabilities. Unless otherwise specified, both stages are trained on 64 GPUs with the AdamW optimizer, gradient clipping set to 1.0, a diffusion timestep shift of 4.0, and a CE:MSE loss weight of 0.25:1 for the generative objectives. Detailed per-task sampling ratios and stage-specific hyperparameters are summarized in Tab. S4.

A.1. Stage 1: LELG Alignment

In the first stage, we train on a mixture of general multimodal understanding datasets—such as FineVision [47] and MAmmoTH [55]—together with our proposed ConsistCompose3M dataset. This joint training retains broad visual–linguistic reasoning capability while injecting layout awareness into both the understanding and generation experts.

To further strengthen multi-instance layout reasoning, we additionally mine high-resolution images (> 512 px) from COCO2017 [21] and Objects365 [35], discarding low-quality samples. These corpora provide dense multi-object scenes (e.g., street environments with pedestrians, cars, bicycles, and traffic lights) as well as structured same-category groupings such as:

3 dogs <bbox>$[x_1, y_1, x_2, y_2]$</bbox>,
<bbox>$[x_1, y_1, x_2, y_2]$</bbox>,
<bbox>$[x_1, y_1, x_2, y_2]$</bbox>

which improve the model's ability to reason about parallel, multi-instance layouts within the LELG framework.

Stage 1 is trained for 18K steps using a cosine learning rate schedule (initial learning rate 2×10^{-5}), no weight decay, and AdamW hyperparameters ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1.0 \times 10^{-15}$). The sequence length per rank ranges from 30K to 38K tokens. For resolution settings, the generation module adopts $(448, 768)$ as its minimum short side and maximum long side, while the understanding module uses $(224, 386)$.

A.2. Stage 2: Hybrid Supervised Fine-Tuning

We further train ConsistCompose on a mixture of generation and understanding tasks, following the hybrid multi-task supervised fine-tuning (SFT) recipe used in Bagel [7]. On the generation side, the mixture comprises text-to-image, multi-reference-to-image, LELG-conditioned text-to-image, LELG-conditioned multi-reference-to-image, and image editing, with layout-conditioned tasks upsampled to reinforce layout consistency. On the understanding side, image-to-text pairs and interleaved understanding data are sampled at lower ratios so as to maintain general visual–language competence while keeping the overall training focused on layout-controllable generation.

Table S4. Training recipe of ConsistCompose.

	LELG Align.	**SFT**
Hyperparameters		
Learning rate	2×10^{-5}	2.0×10^{-5}
LR scheduler	Cosine	Constant
Weight decay	0.0	0.0
Gradient norm clip	1.0	1.0
Optimizer	AdamW	
Loss weight (CE : MSE)	–	0.25 : 1
Warm-up steps	500	500
Training steps	18K	6K
Seq. length / rank (min, max)	(30K, 38K)	(36K, 45K)
Gen res. (short side, long side)	(448, 768)	(512, 1024)
Und res. (short side, long side)	(224, 386)	(224, 386)
Diffusion timestep shift	4.0	4.0
Data sampling ratio		
Text-to-Text	0.05	0.1
Text-to-Image	0.0	0.1
Multi Ref→Image	0.0	0.1
Text-to-Image (LELG)	0.4	0.4
Multi Ref→Image (LELG)	0.4	0.1
Image Editing	0.0	0.1
Image-to-Text pair	0.1	0.1
Interleaved understanding	0.05	0.1

Stage 2 is trained for 6K steps with a constant learning rate of 2.0×10^{-5}, sequence lengths per rank ranging from 36K to 45K tokens, a generation resolution of $(512, 1024)$, and an understanding resolution of $(224, 386)$.

A.3. Data Sampling

Table S4 summarizes the per-task sampling ratios and hyperparameters for both stages. During SFT, layout-conditioned tasks are assigned higher sampling ratios to enhance multi-instance layout controllability, whereas general understanding and interleaved tasks are sampled more sparsely to preserve broad multimodal knowledge.

This two-stage training strategy yields a favorable balance between layout consistency, high-fidelity image generation, and general multimodal understanding.

B. Details of Coordinate-CFG

To improve layout controllability, we introduce a Coordinate-CFG mechanism that augments the text–image classifier-free guidance (CFG) used in Bagel [7] with an additional coordinate-guidance branch. When coordinate

Table S5. List of symbols and their descriptions.

Symbol	Description
$\mathbf{v}_t$	Predicted velocity with full conditioning (text, image, and coordinates)
$\mathbf{v}_t^{\text{text-drop}}$	Predicted velocity without text and coordinate (image-only branch)
$\mathbf{v}_t^{\text{img-drop}}$	Predicted velocity without image
$\mathbf{v}_t^{\text{coord-drop}}$	Predicted velocity without coordinate
$\mathbf{v}_t^{\text{text-cfg}}$	Text-guided velocity in the hierarchical fusion
$\mathbf{v}_t^{\text{coord-cfg}}$	Coordinate-guided velocity in the hierarchical fusion
$\mathbf{v}_t^{\text{final}}$	Final fused velocity after text, coordinate, and image guidance
$s_{\text{text}}, s_{\text{img}}, s_{\text{coord}}$	Guidance scales for text, image, and coordinate branches
$\hat{\mathbf{v}}_t$	Renormalized guided velocity
$\mathcal{N}$	Normalization domain (global or per-channel)

tokens are absent, the formulation reduces exactly to Bagel's original two-branch CFG. Quantitative results on COCO-Position are reported in Table S6.

B.1. Notation

For a noisy latent $\mathbf{x}_t$ at diffusion step t, we denote the velocity predictions of our model as

$$\begin{aligned}
\mathbf{v}_t &= v_\theta(\mathbf{x}_t \mid \text{text}, \text{img}, \text{coord}),\\
\mathbf{v}_t^{\text{text-drop}} &= v_\theta(\mathbf{x}_t \mid \emptyset_{\text{text}}, \text{img}, \emptyset_{\text{coord}}),\\
\mathbf{v}_t^{\text{img-drop}} &= v_\theta(\mathbf{x}_t \mid \text{text}, \emptyset_{\text{img}}, \text{coord}),\\
\mathbf{v}_t^{\text{coord-drop}} &= v_\theta(\mathbf{x}_t \mid \text{text}, \text{img}, \emptyset_{\text{coord}}),
\end{aligned}$$

where v_θ is the velocity prediction network and $\emptyset_{\text{text-drop}}, \emptyset_{\text{img-drop}}, \emptyset_{\text{coord-drop}}$ denote *dropping* (masking out) the text, image, and coordinate conditions, respectively. The superscripts on $\mathbf{v}_t^{\text{text}}, \mathbf{v}_t^{\text{img}}, \mathbf{v}_t^{\text{coord}}$ indicate which modality is *excluded* from the conditioning for that prediction, while the remaining modalities stay active (see the symbol table in Sec. S5).

B.2. Hierarchical fusion

Coordinate-CFG applies classifier-free guidance in a hierarchical manner, extending Bagel's text–image CFG with an additional coordinate-guidance step. The guided velocity is computed as follows:

$$\mathbf{v}_t^{\text{text-cfg}} = \mathbf{v}_t^{\text{text-drop}} + s_{\text{text}}\big(\mathbf{v}_t - \mathbf{v}_t^{\text{text-drop}}\big), \tag{10}$$

$$\mathbf{v}_t^{\text{coord-cfg}} = \mathbf{v}_t^{\text{coord-drop}} + s_{\text{coord}}\big(\mathbf{v}_t^{\text{text-cfg}} - \mathbf{v}_t^{\text{coord-drop}}\big), \tag{11}$$

$$\mathbf{v}_t^{\text{final}} = \mathbf{v}_t^{\text{img-drop}} + s_{\text{img}}\big(\mathbf{v}_t^{\text{coord-cfg}} - \mathbf{v}_t^{\text{img-drop}}\big). \tag{12}$$

This formulation is backward-compatible with the CFG used in Bagel [7]. When the coordinate-guidance branch is disabled at sampling time (i.e., $\mathbf{v}_t^{\text{final}}$ is computed from the text and image branches without the intermediate coordinate-fusion step), the update rule reduces to the original two-branch text–image CFG parameterized by s_{text} and s_{img}. Disabling the image-guidance branch as well further recovers the standard text-only CFG.

B.3. Recommended Coordinate-CFG range

We provide empirically validated coordinate-guidance scales based on MS-Bench [45] and COCO-Position evaluations.

Text-to-Image. A coordinate guidance scale of $s_{\text{coord}} \in [0.6, 3.0]$ offers a favorable balance between enforcing layout constraints and preserving semantic flexibility. Lower values already ensure coarse spatial alignment, while higher values tighten the layout adherence at a moderate cost to sample diversity.

Multi-Reference. For layout-controlled multi-reference generation, we recommend $s_{\text{coord}} \in [0.4, 1.6]$. Within this interval, increasing s_{coord} improves positional stability and identity placement across multiple instances, whereas overly large values tend to over-constrain the layout and degrade visual fidelity.

C. Dataset Construction

We introduce ConsistCompose3M, a large-scale dataset for layout-controllable multi-instance image generation. Compared with existing resources, ConsistCompose3M provides substantially improved *scale*, *quality*, and *adaptability*: it contains millions of diverse multi-instance scenes, includes identity-preserving samples filtered by similarity, and offers structured spatial and semantic supervision suitable for unified multimodal training. The corpus is organized into two complementary partitions—a *layout-grounded text-to-image* split and a *reference-conditioned* split designed for subject-preserving, layout-guided generation. Representative examples of the multi-subject data construction pipeline are shown in Figure S8.

Unified representation. Each instance is annotated with an axis-aligned bounding box and a subject phrase. All spatial annotations are serialized inline using the format `<bbox>[`x_1, y_1, x_2, y_2`]</bbox>`, where (x_1, y_1) and (x_2, y_2) denote the top-left and bottom-right corners of the bounding box. Coordinates are normalized to $[0, 1]$ using the image width W and height H, and are rounded to three decimals. Throughout, we denote an image by I and a normalized bounding box by $b = (x_1, y_1, x_2, y_2) \in [0, 1]^4$.

Primary sources. ConsistCompose3M is constructed from four complementary sources.

Table S6. Quantitative results of the effect of coordinate CFG on the COCO-Position benchmark.

Methods		Instance SR (%) ↑						Image SR (%) ↑						Pos. Acc. (%) ↑			
method	**coord cfg**	L_2	L_3	L_4	L_5	L_6	**Avg**	L_2	L_3	L_4	L_5	L_6	**Avg**	mIoU	AP	AP50	AP75
Bagel		19.1	15.2	13.8	11.4	11.0	13.1	3.7	0.6	0.0	0.0	0.0	0.9	23.1	0.7	3.2	0.2
SFT	0.2	68.4	68.5	62.7	58.3	56.1	61.1	46.3	38.1	21.9	11.9	9.4	25.5	56.0	21.4	45.0	18.4
	0.4	91.9	89.0	87.5	84.6	86.7	87.2	83.8	70.6	60.0	50.0	46.9	62.3	78.4	57.2	82.4	61.3
	0.6	94.7	92.9	90.3	87.0	89.8	90.2	89.4	81.2	66.9	53.7	58.8	70.0	82.3	64.8	86.7	70.2
	0.8	94.7	94.4	91.6	88.4	90.4	91.2	89.4	83.8	70.6	58.8	62.5	73.0	83.6	67.5	87.9	72.8
	1.0	95.3	93.8	91.7	90.2	91.6	92.0	90.6	83.1	70.6	60.6	63.7	73.8	84.4	69.2	88.1	74.4
	1.2	95.3	94.2	92.7	90.2	91.5	92.2	90.6	83.1	71.9	60.6	64.4	74.1	84.8	69.9	89.0	75.4
	1.4	95.0	94.2	93.0	90.6	92.4	92.6	90.6	83.8	73.1	62.5	68.8	75.7	85.2	70.5	89.1	75.5
	1.6	95.6	94.2	92.7	90.6	92.4	92.6	91.9	83.1	73.1	63.7	68.8	76.1	85.3	70.9	89.1	76.9
	1.8	96.3	94.4	92.5	91.2	92.8	92.9	93.1	83.8	72.5	65.0	70.0	76.9	85.6	71.7	89.4	77.4
	2.0	96.6	94.6	93.4	90.7	92.1	92.8	93.1	85.0	75.6	64.4	66.9	77.0	85.6	71.2	88.8	78.1
	2.2	95.9	95.2	93.3	90.9	92.5	93.0	92.5	86.3	75.6	62.5	68.8	77.1	85.5	71.2	89.2	76.8
	2.4	96.3	94.6	93.1	89.7	92.3	92.6	92.5	85.0	75.0	61.9	65.0	75.9	85.4	71.2	88.9	76.8
	2.6	96.3	94.6	92.5	89.7	91.9	92.3	92.5	85.0	73.8	60.6	65.6	75.5	85.2	70.9	89.1	76.7
	2.8	96.6	95.2	93.1	89.2	92.1	92.5	93.1	86.9	75.6	60.6	66.9	76.6	85.2	70.9	89.1	77.1
	3.0	96.9	94.8	93.4	90.6	91.6	92.7	93.8	85.6	76.9	63.1	65.0	76.9	85.4	71.1	89.3	77.1

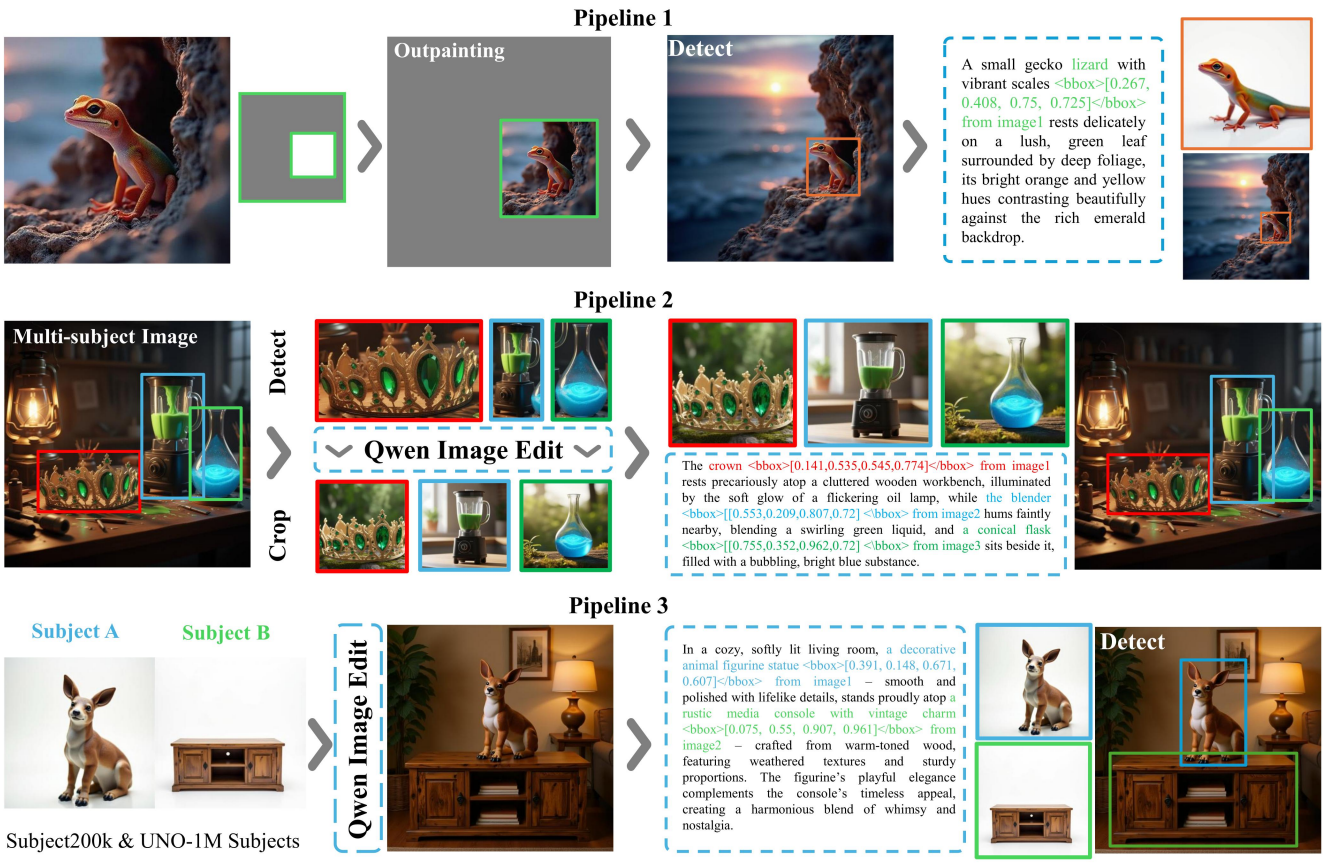

Figure S8. Overview of the reference-conditioned branch in ConsistCompose3M construction, comprising three pipelines. All annotations are represented using a unified inline <bbox> format.

(1) LayoutSAM. LayoutSAM [57] provides multi-object scenes with instance-level bounding boxes and textual descriptions. We convert its annotations into our unified training format and, for the layout-grounded text-to-image split, additionally supply a global scene description for each image that summarizes all entities and their relations.

(2) UNO-1M. UNO-1M [51] is built on top of Objects365 [35] and derives subject-centric data from its category set. We follow the class taxonomy used in UNO-1M [51]and directly reuse its processed outputs—subject reference images and attribute-aware phrases—as paired sets $\{R_i\}$ and $\{s_i\}$. No additional subject categories are introduced.

(3) Subjects200K. Subjects200K [37] is incorporated in the same manner, expanding both the coverage and appearance diversity of the subject bank.

(4) Virtual try-on data. To further enrich the reference-conditioned branch, we incorporate human–garment virtual try-on datasets such as VITON-HD [5] and DressCode-MR [6]. Each sample consists of a reference garment image and a corresponding try-on image in which the garment is embedded within a layered human-body configuration, naturally inducing nested and occluded spatial relationships. We treat the garment image as the reference, annotate an axis-aligned bounding box for the garment region in the try-on image, and describe the scene using a standardized English virtual-try-on template.these samples provide complementary reference-conditioned supervision in which compositional constraints arise from body–garment interactions rather than explicit multi-object layouts.

C.1. Two data construction branches

We construct two complementary data branches that share the same inline `<bbox>` representation but differ in whether reference images are provided.

C.1.1. Layout-grounded T2I branch.

This branch is built from two sources: (i) LayoutSAM-style corpora [57] with given images, a *global dense caption*, and instance-level spatial annotations, and (ii) canonicalized samples derived from our reference-conditioned pipelines. For (i), LayoutSAM [57] provides, for each image, a global dense caption $p_{\text{dense}}^{\text{raw}}$ and a set of entities with class labels and bounding boxes, where each instance is already mentioned in $p_{\text{dense}}^{\text{raw}}$.

We convert the global dense caption $p_{\text{dense}}^{\text{raw}}$ to a layout-grounded LELG prompt p_{LELG}, preserving its narrative content while explicitly encoding all annotated instances and their spatial coordinates, using GPT-4o in an in-context learning configuration with the prompt design shown in Fig. S12. For each instance, we append its normalized bounding-box tag `<bbox>`$[x_1, y_1, x_2, y_2]$`</bbox>` immediately after the first occurrence of the corresponding phrase in p_{LELG}, producing layout-grounded T2I training pairs while transforming dense spatial annotations into inline, language-embedded supervision.

For (ii), we start from the outputs of Pipelines 1–3 (described below) and apply the deterministic conversion in Sec. C.1.2: we drop all reference-image inputs and remove any "from image k" indices from the tags, while keeping the scene sentence otherwise unchanged.

C.1.2. Reference-conditioned branch

This branch is built from UNO-1M [51] and Subjects200K [37] subject banks and phrases processed by the three pipelines below (see Fig. S8). Each sample consists of a scene sentence with inline `<bbox>` tags and a set of subject reference images $\{R_i\}$. The pipelines are explicitly designed for *subject-consistent, layout-guided generation*. Through the canonicalization in Sec. C.1.1, the same samples also contribute to the layout-grounded T2I branch. In addition, we use virtual try-on data from VITON-HD [5] and DressCode-MR[6] as auxiliary sources, converting them into the same reference-conditioned format to further enrich the subject and appearance diversity of ConsistCompose3M.

Pipeline 1: Layout Diversification via Controlled Outpainting. As shown in Pipeline 1 of Fig. S8, this pipeline broadens the spatial distribution of bounding-box positions and scales in the single-reference setting using image pairs from Subjects200K [37], obtained through OmniControl's side-by-side paired generation. For each sample, we first generate a dense caption $p_{\text{dense}}^{\text{raw}}$ for the reference image using GPT-4o [14]. The reference image R_i is then resized by a sampled side-length factor $r \sim \mathcal{U}(0.6, 0.8)$ and placed onto a blank canvas at a random valid location, thereby defining a normalized bounding box $b \in [0, 1]^4$.

FLUX.1 Fill is subsequently applied to outpaint the surrounding regions, producing an initial scene completion while preserving the pasted reference. A feathered binary mask is finally used to blend the pasted region with the outpainted context, yielding a coherent image I. Because subjects in Subjects200K reference image are typically centered, the random rescaling and repositioning effectively counteract this center bias and support a wider range of spatial configurations. The resulting box b is recorded as the inline `<bbox>` annotation, providing diverse layouts and informative supervision for layout-aware generation.

Pipeline 2: Crop-based single-reference refinement. As illustrated in Pipeline 2 of Fig. S8, this pipeline refines single-reference samples by extracting and enhancing subject crops from multi-subject scenes. We begin by applying GroundingDINO to detect subject-specific bounding boxes, retaining boxes with area ratios in $[0.20, 0.60]$, removing high-IoU duplicates, and requiring at least three valid subjects per image. For each surviving subject–box pair (R_i, b), we extract the corresponding crop $I|_b$ and retain candidates that satisfy the CLIP-T constraint $c(I|_b, R_i) \geq 0.25$. The selected crop is then refined using Qwen Image Edit as shown in Fig. S9, yielding a generated image I'. We subsequently filter the generation by comparing I' directly with the original crop $I|_b$ using CLIP-I/CLIP-T scores, and

Figure S9. Crop-based single-reference refinement. Starting from a multi-subject scene, a subject crop is extracted using its detected bounding box and refined by Qwen Image Edit in a subject-to-image manner, producing an identity-preserving single-reference sample.

convert the resulting valid candidate into an inline `<bbox>` tag via ICBP.

Pipeline 3: Multi-Reference Guided Scene Synthesis. As shown in Pipeline 3 of Fig. S8, we construct multi-subject scenes by sampling $K \in \{2, \ldots, 4\}$ distinct subject–reference pairs $\{(s_i, R_i)\}$ from UNO-1M (Object365-derived) and Subjects200K. We then query GPT-4o with the instruction template in Fig. S13 to obtain a concise global dense prompt $p^{\text{raw}}_{\text{dense}}$ that includes each subject phrase s_i verbatim within the description.

Qwen Image Edit synthesizes a multi-subject image I from the image references $\{R_i\}$ and the text prompt $p^{\text{raw}}_{\text{dense}}$. Instead of directly pasting reference crops onto a canvas, the model re-renders the subjects coherently within the scene, substantially reducing trivial copy–paste artifacts commonly observed in multi-reference generation.

To link each subject reference to a concrete location in I and filter out erroneous generations, we run GroundingDINO on I for each subject phrase to obtain subject-specific candidate boxes $\{b_j\}$. For every subject s_i with its reference image R_i and every candidate box b_j, we crop the corresponding region $I|_{b_j}$ and compute three scores: (i) a CLIP text–image similarity s^{T}_{ij} between $I|_{b_j}$ and the subject phrase s_i, (ii) a CLIP image–image similarity s^{I}_{ij} between $I|_{b_j}$ and the reference image R_i, and (iii) a GroundingDINO detection score s^{D}_{ij} associated with b_j. After normalizing all scores to $[0, 1]$, we form a combined similarity

$$S_{ij} = \alpha\, s^{\text{T}}_{ij} + \beta\, s^{\text{D}}_{ij} + \gamma\, s^{\text{I}}_{ij}, \qquad C_{ij} = 1 - S_{ij}, \tag{13}$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$, and C_{ij} is the cost of assigning subject i to box j. We then solve the linear assignment problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \sum_{i=1}^{M} \sum_{j=1}^{N} C_{ij} X_{ij} \\ \text{s.t.} \quad & \mathbf{X} \in \{0,1\}^{M \times N}, \\ & \sum_{j=1}^{N} X_{ij} = 1, \quad i = 1, \ldots, M, \\ & \sum_{i=1}^{M} X_{ij} \leq 1, \quad j = 1, \ldots, N. \end{aligned} \tag{14}$$

via the Hungarian algorithm, where M and N denote the numbers of subjects and detected boxes, respectively, and $\mathbf{X}$ is the assignment matrix. This yields a unique bounding box in I for each subject, or rejects the scene if a complete, non-duplicated matching is impossible. We further discard a scene if any matched pair fails per-pair quality checks (e.g., CLIP-T, CLIP-I, or detection scores falling below preset thresholds), so this matching step simultaneously aligns references to spatial positions in the synthesized image and aggressively filters out low-quality or mismatched generations.

Finally, the three pipelines collectively yield, for each constructed sample, a scene-level dense description and the corresponding subject–bounding-box associations. We then convert the raw dense sentence $p^{\text{raw}}_{\text{dense}}$ into a layout-grounded, multi-reference prompt p_{LELG} by inserting explicit bounding-box and source-image tags. For each matched subject, we take its normalized bounding box $[x_1, y_1, x_2, y_2]$ together with the index of its reference image and employ GPT-4o [14] with the instruction template in Fig. S14 to insert a tag of the form `<bbox>[`x_1, y_1, x_2, y_2`]</bbox> from image` k immediately after the first noun-phrase mention of that subject in $p^{\text{raw}}_{\text{dense}}$. By integrating reference-guided synthesis, detection-and-matching–based filtering, and LLM-driven tag insertion, these pipelines produce layout-grounded samples exhibiting rich nested and attributive configurations.

D. Supplementary Results and Detailed Analysis

Layout-Grounded Text-to-Image Generation. We provide additional qualitative results to complement the main paper. Fig. S15 presents extended examples of layout-grounded multi-instance generation under the ICBP paradigm. Moreover, Fig. S16 showcases more challenging,

Table S7. Quantitative results on GenEval. * indicates methods using an LLM-based rewriter.

Model	Single	Two	Count	Color	Pos.	Attr.	Overall↑
SDXL	0.98	0.74	0.39	0.85	0.15	0.23	0.55
FLUX.1-dev	0.99	0.82	0.78	0.73	0.19	0.46	0.66
Show-o	0.98	0.80	0.67	0.83	0.32	0.50	0.68
Janus	0.97	0.68	0.31	0.84	0.46	0.42	0.61
Janus Pro	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Bagel*	0.98	0.95	0.84	0.93	0.72	0.73	0.86
Alignment*	0.98	0.97	0.77	0.93	0.84	0.80	0.88
SFT*	0.98	0.95	0.79	0.94	0.84	0.79	0.88

Table S8. Editing performance on GEdit-Bench-EN.

Model	SC↑	PQ↑	OS↑
Bagel Base	6.94	6.73	6.36
Ours (w/o Coord Data)	6.09	6.72	5.87
Ours (w/ Coord Data)	6.32	6.82	5.78

heavily constrained scenarios with overlapping and closely interacting subjects, where our method continues to preserve object count, fine-grained attributes (e.g., color and pose), and precise spatial arrangements. Across these cases, the model maintains high visual fidelity and a consistent global style, further substantiating the quantitative improvements.

General Capabilities. We study whether introducing ICBP coordinate data for layout grounding affects the model's general capabilities in generation, image editing, and multimodal understanding.

For generation, we evaluate fine-grained control on GenEval (Table S7). Our Alignment* and SFT* (with an LLM-based rewriter) achieve an overall score of 0.88, with strong position control (0.84) and attribute consistency (0.79–0.80), indicating that precise layout adherence does not compromise control granularity.

For editing, to control for data/recipe confounders, we evaluate on GEdit-Bench-EN (Table S8) with three settings: (i) *Bagel Base*, the original pretrained model; (ii) *Ours (w/o Coord Data)*, trained with the same data mixture and optimization recipe but without injecting ICBP coordinate data; and (iii) *Ours (w/ Coord Data)*, trained identically while including such coordinate data. Both trained variants show a mild drop compared to Bagel Base, while *w/ Coord Data* is comparable to *w/o Coord Data* under matched data and optimization, suggesting that ICBP coordinate data does not introduce additional regression on general editing.

For understanding, we report results on general multimodal benchmarks (Table S9). Alignment variants (with/without coord data) achieve ∼81.4 on MMBench, comparable to InternVL3-8B (81.7) [64] and Bagel (81.4),

Table S9. Quantitative results on general understanding benchmarks.

Model	MMBench↑	MMMU↑
Qwen2.5-VL-3B-Instruct	77.4	53.1
Qwen2.5-VL-7B-Instruct	82.6	58.6
InternVL3-8B	81.7	65.6
Bagel	81.4	46.4
Alignment (w/o Coord Data)	81.5	39.4
Alignment (w/ Coord Data)	81.4	42.3
SFT (w/ Coord Data)	75.3	40.2

Table S10. Quantitative results on DreamBench (Single/Multi)

Method	DreamBench Single			DreamBench Multi		
	DINO	CLIP-I	CLIP-T	DINO	CLIP-I	CLIP-T
Step1X-Editing	0.616	0.779	0.314	–	–	–
OmniGen	0.554	0.746	0.322	0.441	0.692	0.341
OmniGen2	0.671	0.791	0.312	0.459	0.698	0.333
UNO	0.661	0.796	0.304	0.491	0.715	0.323
Flux Kontext dev	0.687	0.806	0.310	0.500	0.712	0.328
Qwen Image Edit	0.638	0.780	0.326	0.458	0.697	0.342
GPT4O	0.687	0.801	0.310	0.529	0.725	0.324
Alignment (w/o Coord Data)	0.565	0.757	0.325	0.456	0.699	0.342
Alignment (w/ Coord Data)	0.611	0.770	0.320	0.473	0.701	0.337
SFT (w/ Coord Data)	0.677	0.792	0.314	0.506	0.703	0.335

Table S11. COCO-Position: comparison with general image generation models. Instance/Image success ratio (Avg, %) and position accuracy (%).

Method	Inst. SR	Img. SR	mIoU	AP	AP50	AP75
QwenImage	7.9	0.3	18.6	0.3	1.5	0.1
Nano Banana	11.2	0.1	22.8	0.5	2.4	0.1
Ours	92.6	76.1	85.3	70.9	89.1	76.9

validating no degradation in basic visual understanding. Notably, on MMMU, introducing coord data improves Alignment from 39.4 (w/o coord data) to 42.3 (w/ coord data), suggesting coord data may benefit spatial reasoning.

Extended DreamBench Results. To assess generalization beyond layout control, we additionally report DreamBench results (Tab. S10, Figs. S17 and S18). *Alignment (w/ Coord Data)* outperforms its non-coordinate counterpart (0.611 vs. 0.565 in the DreamBench single-subject DINO identity metric), and *SFT (w/ Coord Data)* further reaches 0.677 for single-subject and 0.506 for multi-subject generation, surpassing baselines such as OmniGen2 [50] and narrowing the gap to top-performing models (Flux Kontext dev [17], GPT-4o [14]). Qualitative examples in Figs. S17 and S18 corroborate these gains, showing strong identity preservation and prompt adherence for single subjects, as well as consistent identities with natural interactions in multi-subject scenes. Together, these results indicate that our method attains strong layout control without compromising general understanding or broad generative capability.

Comparison with general image generation models. We compare our model with general-purpose image generation models on COCO-Position (Table S11). Despite strong text-to-image quality, these general models struggle to satisfy explicit multi-instance spatial constraints, resulting in very low instance/image success rates and poor localization accuracy. In contrast, our unified model achieves substantially higher success rates and more accurate localization across all metrics.

Figure S10. Interaction inconsistency & unintended additions.

Failure Case Analysis. As illustrated in Figure S10, failures arise when the model cannot infer a geometry-consistent global pose and occlusion/contact topology from the layout. This leads to two typical error patterns: (a) *inconsistent interactions* between objects, and (b) *unintended additions* that only satisfy local spatial constraints while violating global consistency.

Requirements

1. When a subject appears multiple times in the original prompt:
(a) If referring to the same instance, merge mentions and place tags after the first occurrence.
(b) If referring to different instance, tag each occurrence separately.
(c) Add count prefixes (e.g., "`2 backpacks`") when appropriate for clarity.

2. For subjects with single bbox: place the tag immediately after the first occurrence of the subject, e.g., `[subject] <bbox>[...]</bbox>`.

3. For subjects with multiple bboxes: ensure the count matches the number of bboxes provided, e.g., `[count] [subject] <bbox>[...]</bbox>, ..., <bbox>[...]</bbox>`.

4. Make minimal necessary modifications only to:
(a) Improve clarity when adding multiple bbox tags.
(b) Resolve redundancy in repeated subject mentions.
(c) Adjust grammar for coherence after tag insertion.

5. Preserve all original information. including:
(a) Specific details and attributes.
(b) Overall meaning and context.
(c) Sentence structure when possible.

6. Output Format (Critical): Only return the modified prompt with bbox tags. Do *not* add any extra content, such as titles (e.g., "Modified Prompt"), comments, or explanations.

Example 1
Original: "A hiker wears a backpack. The backpack is black. He also carries a water bottle."
Tags: "backpack" (Details: a large black hiking backpack) → `1 backpack <bbox>[0.3,0.4,0.5,0.7]</bbox>`
Expected Output: "A hiker wears a backpack <bbox>[0.3,0.4,0.5,0.7]</bbox> which is black. He also carries a water bottle."

Example 2
Original: "Dogs play in the park. The dogs chase each other happily."
Tags: "dogs" → `3 dogs <bbox>[0.1,0.2,0.3,0.4]</bbox>, <bbox>[0.5,0.6,0.7,0.8]</bbox>, <bbox>[0.2,0.3,0.4,0.5]</bbox>`
Expected Output: "3 dogs <bbox>[0.1, 0.2, 0.3, 0.4]</bbox>, <bbox>[0.5, 0.6, 0.7, 0.8]</bbox>, <bbox>[0.2, 0.3, 0.4, 0.5]</bbox> play in the park. The dogs chase each other happily."

Figure S11. Requirements used in the prompt template for layout-grounded T2I branch. This part specifies how to insert inline `<bbox>` tags into a dense caption $p_{\text{dense}}^{\text{raw}}$.

Global Dense Caption:
This is a photo showcasing a corner of a city street, with a row of **orange-red buildings** in the background. In the foreground, there is a green trash can with a **vintage-style advertisement** on it, featuring a man and a woman in a classic movie poster style. Next to the trash can is a **wooden bench**, with a sign displaying a schedule of performances on a stand in front of it. The entire scene is illuminated by natural light, creating a tranquil and nostalgic atmosphere.

Instance with Spatial Annotation:
`'wooden bench'` → `<bbox>[0.497,0.725,0.998,0.995]</bbox>`
`'vintage-style advertisement'` → `<bbox>[0.2,0.326,0.386,0.813]</bbox>`
`'orange-red buildings'` → `<bbox>[0.002,0.003,0.997,0.876]</bbox>`

Requirements:
(omitted here for brevity; see main prompt in Fig. S11)

Resulting LELG prompt:

This is a photo showcasing a corner of a city street, with a row of orange-red buildings <bbox>[0.002,0.003,0.997,0.876]</bbox> in the background. In the foreground, there is a green trash can with a vintage-style advertisement <bbox>[0.2,0.326,0.386,0.813]</bbox> on it, featuring a man and a woman in a classic movie poster style. Next to the trash can is a wooden bench <bbox>[0.497,0.725,0.998,0.995]</bbox>, with a sign displaying a schedule of performances on a stand in front of it. The entire scene is illuminated by natural light, creating a tranquil and nostalgic atmospher.

Figure S12. Prompt template for layout-grounded T2I branch. The figure shows the input dense caption $p_{\text{dense}}^{\text{raw}}$, its instance–bbox annotations, and the resulting layout-grounded sentence p_{LELG} with inline `<bbox>` tags.

Create a detailed English scene description that includes the following exact subjects:
{subject list}

IMPORTANT REQUIREMENTS:

1. Use the exact subject names as provided — *do not* paraphrase or rename them.
2. Show clear interactions between the subjects.
3. Add specific environment details (lighting, setting, time of day).
4. Use vivid, descriptive language suitable for image generation.
5. Keep it concise (1–2 sentences); no explanations, just the scene.

Example:
`["cat", "yarn ball"]`
Example output:
"The cat plays with the yarn ball on a sunny windowsill, batting at the colorful strands with its paw."

Figure S13. Prompt template for generating $p_{\text{dense}}^{\text{raw}}$ (Reference-conditioned Branch)

Your task is to modify the original English prompt by inserting a bbox tag **after** each specified subject, **without** changing any other content of the original prompt.

Original:
{original_prompt}

Tags (keep original names, insert tag immediately after the subject):
{Subjects to tag }

Requirements:

1. Do not add, delete, or paraphrase any words in the original prompt except inserting the bbox tags.
2. Place the bbox tag immediately after the exact subject (e.g., "a boy" → "a boy<bbox>[x1,y1,x2,y2]</bbox> from image N").
3. Return only the modified prompt (no extra explanations or sentences).
4. Insert the tag only after the **first** occurrence of each subject (if duplicated).

Example:
Original:
"The cat plays with the yarn ball on a sunny windowsill."
Tags:
"cat" → `<bbox>[0.1,0.2,0.3,0.4]</bbox>` from image1;
"yarn ball" → `<bbox>[0.5,0.6,0.7,0.8]</bbox>` from image2
Modified:
"The cat<bbox>[0.1,0.2,0.3,0.4]</bbox> from image1 plays with the yarn ball<bbox>[0.5,0.6,0.7,0.8]</bbox> from image2 on a sunny windowsill."

Figure S14. Prompt template for inserting `<bbox>` tags into $p_{\text{dense}}^{\text{raw}}$ (Reference-conditioned Branch)

A tiny, determined puppy <bbox>[0.221, 0.208, 0.773, 0.89]</bbox> paddles gracefully in crystal-clear waters <bbox>[0.003, 0.006, 0.996, 0.996]</bbox>. Its eyes gaze forward, ears perked and fur glistening in sunlight. Small ripples form beneath its paws, and soft reflections add calm—its curiosity and spirit shine in this moment of exploration.

CG rendering of a dinosaur-steampunk creature <bbox>[0.037, 0.064, 0.916, 0.962]</bbox>. Wide-open mouth <bbox>[0.49, 0.326, 0.91, 0.707]</bbox> shows sharp teeth, prominent eye <bbox>[0.451, 0.23, 0.56, 0.367]</bbox> encircled by metallic gear-like structures. Blue-white skin with rust patches blends organic-mechanical worlds, set against blurred backdrop highlighting bold design.

Central artwork depicts a surreal steam locomotive <bbox>[0.005, 0.008, 0.961, 0.977]</bbox> with a man's face integrated into its design. The man's exaggerated stern expression <bbox>[0.484, 0.180, 0.751, 0.518]</bbox> takes center stage at the front. Billowing steam <bbox>[0.003, 0.748, 0.397, 0.967]</bbox> envelops the background, blending fantasy and machinery into a vivid, striking visual.

In a shadowy, mysterious setting, there are two intriguing figures that draw the eye. On the left, a whimsical character with an oversized, round head and tiny wings stands with an expression of bewilderment<bbox>[0.076, 0.247, 0.443, 0.891]</bbox>. This figure appears to be part of a larger artistic composition. To the right, an ornate sculpture with swirling patterns houses a somewhat melancholy character, embedded as part of the artwork <bbox>[0.491, 0.069, 0.923, 0.902]</bbox>. A significant element between these two figures is a skull, its hollow eyes and simple features capturing an eerie charm <bbox>[0.541, 0.246, 0.811, 0.574]</bbox>. Below this, a small zombie-like figure emerges, its posture slightly bent as if attempting to communicate with its larger companion <bbox>[0.587, 0.557, 0.750, 0.733]</bbox>. The scene is set against a dark, rocky background that adds to the intriguing, otherworldly atmosphere.

In a shadowy, mysterious setting, there are two intriguing figures that draw the eye. On the left, a whimsical character with an oversized, round head and tiny wings stands with an expression of bewilderment<bbox>[0.076, 0.247, 0.443, 0.891]</bbox>. This figure appears to be part of a larger artistic composition. To the right, an ornate sculpture with swirling patterns houses a somewhat melancholy character, embedded as part of the artwork <bbox>[0.491, 0.069, 0.923, 0.902]</bbox>. A significant element between these two figures is a skull, its hollow eyes and simple features capturing an eerie charm <bbox>[0.541, 0.246, 0.811, 0.574]</bbox>. Below this, a small zombie-like figure emerges, its posture slightly bent as if attempting to communicate with its larger companion <bbox>[0.587, 0.557, 0.750, 0.733]</bbox>. The scene is set against a dark, rocky background that adds to the intriguing, otherworldly atmosphere.

In this captivating scene, a majestic animal <bbox>[0.206, 0.078, 0.680, 0.524]</bbox>, its fur gleaming in the low light, peers intently into a large, ornate pot <bbox>[0.192, 0.434, 0.794, 0.901]</bbox> that seems to shimmer with mysterious energy. This pot rests atop what appears to be a well-worn caldron <bbox>[0.008, 0.618, 0.212, 0.902]</bbox>, its surface adorned with ornate patterns that suggest a long history. Rising dramatically from the mix of these intriguing elements is a swirling steam <bbox>[0.685, 0.005, 0.928, 0.896]</bbox>, curling and twisting into the air, giving the scene an otherworldly, almost mystical ambiance. The combination of these subjects creates a striking visual narrative that feels both ancient and powerful.

Figure S15. Qualitative examples of ConsistCompose for layout-controllable multi-instance generation under the ICBP paradigm. The model performs text-driven T2I generation with faithful prompt alignment and high visual fidelity, and synthesizes subjects with precise attributes and spatial arrangements.

A cozy study room scene, a cactus <bbox>[0.55,0.198,0.79,0.413]</bbox> with golden spines sits in a tall blue flower pot <bbox>[0.557,0.41,0.798, 0.643]</bbox> on top of a stack of book <bbox>[0.56,0.64,0.827,0.68]</bbox>, book <bbox>[0.548,0.673,0.827, 0.73]</bbox>, book <bbox>[0.563, 0.725, 0.872, 0.778]</bbox> and book <bbox>[0.578, 0.768, 0.773, 0.843]</bbox>, while on the left a tall branching cactus <bbox>[0.188, 0.078, 0.435, 0.548]</bbox> rises behind a terracotta flower pot <bbox>[0.19, 0.54, 0.435, 0.655]</bbox> filled with three upright columnar cactus <bbox>[0.218, 0.423, 0.287, 0.547]</bbox>,cactus <bbox>[0.282,0.42,0.343,0.553]</bbox> and cactus <bbox>[0.343,0.42,0.42,0.555]</bbox> resting on a neat pile of book <bbox>[0.147, 0.633, 0.437, 0.688]</bbox>, book <bbox>[0.148,0.68,0.423,0.723]</bbox> and book <bbox>[0.155,0.713,0.408,0.763]</bbox>, in front of which a low white flower pot <bbox>[0.338,0.747,0.61,0.83]</bbox> holds a tall slender cactus <bbox>[0.433,0.462,0.5,0.758]</bbox>, a slightly shorter cactus <bbox>[0.497,0.547,0.548,0.77]</bbox> and a small round cactus <bbox>[0.377,0.678,0.498,0.767]</bbox> perched on two more book <bbox>[0.24,0.79,0.513,0.867]</bbox> and book <bbox>[0.262,0.877,0.513,0.91]</bbox>, with an additional horizontal book <bbox>[0.525,0.832,0.74,0.923]</bbox> supporting the right stack and smooth stone <bbox>[0.06,0.743,0.26,0.94]</bbox> and stone <bbox>[0.737,0.788,0.955,0.923]</bbox> placed at the front edges of the arrangement.

Figure S16. Layout-grounded text-to-image generation in an extremely dense multi-instance scene with overlapping and nested subjects.

Figure S17. Qualitative comparison on DreamBench single-subject generation. Across diverse prompts, ConsistCompose preserves subject identity and follows the textual description comparably to state-of-the-art methods.

Figure S18. Qualitative comparison on DreamBench multi-subject generation. In more challenging multi-subject and multi-scenario settings, ConsistCompose maintains consistent subject appearance and prompt faithfulness on par with state-of-the-art methods.