Training-Free Spatial Control for Multi-Entity Text-to-Image Generation

Bounding-Box Adherence for Complex Compositional Prompts

MSc DSAIT Thesis Project

Vladimir Petkov

Delft University of Technology

TU Delft

Training-Free Spatial Control for Multi-Entity Text-to-Image Generation

Bounding-Box Adherence for Complex Compositional Prompts

by

Vladimir Petkov

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday, June 19, 2026, at 4:00 PM.

Student number:	5447194	
Project duration:	August 20, 2025 – June 19, 2026	
Thesis committee:	Dr. ir. H. Jamali-Rad,	TU Delft, Thesis Daily Supervisor
	Dr. ir. E. Isufi,	TU Delft, Thesis Advisor
	Dr. ir. H. Palangi,	Google Research, External Co-supervisor
	Dr. ir. J. Martinez Castaneda,	TU Delft, Committee Chair
	Dr. ir. M. Skrodzki,	TU Delft, External Committee Member

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

This thesis concludes my MSc in Data Science and Artificial Intelligence Technology at Delft University of Technology.

I am deeply grateful to Dr. Hadi Jamali-Rad for his advising throughout this project: for always pushing me to strive for the best work I could do; for his close involvement and communication; and for the feedback that sharpened both the work and the way I think about research. I would also like to thank Dr. Hamid Palangi and Dr. Tejas Gokhale for their collaboration and the external expertise they brought, raising questions and ideas that had never crossed my mind. The weekly calls where I presented my progress to them sharpened not only the research but also the way I communicate it.

I owe special thanks to Nhat Vy Dinh, Athanasios Masouris, and Sieger Falkena, who stepped in during the final weeks to help bring this work to its finished form. Their help with the evaluation runs and the writing of the WACV submission made it possible to get everything into good shape, and I am grateful for their generosity at exactly the moment it was needed most.

Finally, I want to thank my friends and my family for their patience and support over the past year, especially through the weeks when the results refused to cooperate. I hope you enjoy reading this thesis.

Vladimir Petkov
Delft, June 2026

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
CFG	Classifier-Free Guidance
CFM	Conditional Flow Matching
DiT	Diffusion Transformer
DWC	Distance-Weighted Containment
FM	Flow Matching
FOCAL	Flow Optimal Control for Alignment in Layout-Guided T2I Diffusion
ICL	In-Context Learning
JSD	Jensen–Shannon Divergence
KL	Kullback–Leibler (divergence)
LLM	Large Language Model
MM-DiT	Multi-Modal Diffusion Transformer
ODE	Ordinary Differential Equation
RF	Rectified Flow
SDE	Stochastic Differential Equation
SGE	SpatialGenEval (and the curated SGE-mini benchmark)
SOC	Stochastic Optimal Control
SOTA	State-of-the-Art
T2I	Text-to-Image
VAE	Variational Autoencoder
VLM	Vision–Language Model
VQA	Visual Question Answering

Symbols

Symbol	Definition	Domain
$\mathbf{A}$, $\mathbf{A}^{(k)}$	Image-to-text joint-attention readout, averaged over blocks / for block k	$\mathbb{R}^{N_{\text{img}} \times N_{\text{txt}}}$
$a(t)$	Costate (adjoint variable), $a(t) = \nabla_x V$	$\mathbb{R}^d$
b	Flow-matching base drift of the controlled SDE	$\mathbb{R}^d$
B_t	Standard Brownian motion	$\mathbb{R}^d$
$\mathcal{B}_e$	Bounding box of entity e	$\subseteq [0,1]^2$
c	Text conditioning (prompt)	text
$\mathbf{c}_e$	Centre of bounding box $\mathcal{B}_e$	$[0,1]^2$
C, H, W	Latent channel count, height, width	$\mathbb{N}$
d	Latent dimension, $d = CHW$	$\mathbb{N}$
$d_e(\mathbf{p})$	ℓ_∞ exterior distance from position $\mathbf{p}$ to $\mathcal{B}_e$	$\mathbb{R}_{\geq 0}$
$\mathcal{D}$	VAE decoder	operator
E	Number of entities in the prompt	$\mathbb{N}$
$\mathcal{E}$	VAE encoder	operator
$\mathcal{E}(c)$	Entity set of the scene graph $\mathcal{G}(c)$	set
f	Running cost	$\mathbb{R}$
$\mathcal{G}(c)$	Scene graph of prompt c, $\mathcal{G}(c) = (\mathcal{E}(c), \mathcal{R}(c))$	graph

Symbol	Definition	Domain
g	Terminal cost ($g \equiv 0$ in FOCAL)	$\mathbb{R}$
$\mathcal{H}$	Hamiltonian	$\mathbb{R}$
$\mathbf{I}$	Identity matrix	$\mathbb{R}^{d\times d}$
$J(u)$	Control cost function	$\mathbb{R}$
K	Number of joint-attention blocks	$\mathbb{N}$
L_d	Disentanglement loss	$\mathbb{R}_{\geq 0}$
L_{coh}, L_{sep}	Coherence / separation sub-terms of L_d	$\mathbb{R}_{\geq 0}$
L_t	Centroid-translation loss	$\mathbb{R}_{\geq 0}$
L_c	Distance-weighted containment loss	$\mathbb{R}_{\geq 0}$
N	Number of Euler sampling steps	$\mathbb{N}$
N_{img}, N_{txt}	Number of image / text tokens	$\mathbb{N}$
$\mathbf{p}$	Image position, $\mathbf{p} = (p_x, p_y)$	$[0,1]^2$
p_t	Marginal density of the state X_t at time t	density
$\mathbf{Q}$	Stacked per-token attention maps, $\mathbf{Q}_j = \mu_j$	$\mathbb{R}^{T\times N_{\text{img}}}$
$\mathcal{R}(c)$	Relation (triplet) set of the scene graph	set
s	Classifier-free-guidance scale	≥ 1
$\mathcal{T}_e$	Token group (indices) of entity e	index set
t	Sampling time (t=0 noise, t=1 data)	$[0,1]$
u, $u^\star$	SOC control / optimal control	$\mathbb{R}^d$
$V(x,t)$	Value function	$\mathbb{R}$
v_θ	Trained (CFG) velocity field	$\mathbb{R}^d$
$\tilde{v}_\theta$	FOCAL-corrected velocity	$\mathbb{R}^d$
v_t^u	Controlled velocity	$\mathbb{R}^d$
w_d, w_t, w_c	Cost-term weights (disentangle, translate, contain)	$\mathbb{R}_{\geq 0}$
$w(t)$	Velocity-correction weight, $w(t) = \lambda\,\eta(t)$	$\mathbb{R}_{\geq 0}$
X_t	Latent state at time t (X_0 noise, X_1 data)	$\mathbb{R}^d$
X_t^u	Controlled latent state	$\mathbb{R}^d$
$\bar{X}_t$	Reference-path state	$\mathbb{R}^d$
$\mathbf{X}^{(\text{img})}$, $\mathbf{X}^{(\text{txt})}$	Image / text token sequences	$\mathbb{R}^{N\times D}$
x, y, w, h	Layout box: top-left x, y, width w, height h (normalised)	$[0,1]$
α_t, β_t	Flow-matching interpolation schedules; RF: $(\alpha_t, \beta_t) = (t, 1-t)$	$[0,1]$
$\delta(q;a,b)$	Per-axis exterior distance of coordinate q from $[a,b]$	$\mathbb{R}_{\geq 0}$
$\eta(t)$	SOC step-size weight $\frac{1}{2}\sigma^2(t)(1-t)$; memoryless $(1-t)^2/t$	$\mathbb{R}_{\geq 0}$
θ	Network (MM-DiT) parameters	—
λ	Global guidance strength (scalar gain)	$\mathbb{R}_{\geq 0}$
μ_j	Per-token spatial attention map	simplex
$\bar{\mu}_e$	Per-entity attention map	simplex
π_0, π_1, π_t	Noise, data, and intermediate marginal distributions	dist.
$\sigma(t)$	Diffusion schedule; $\sigma_{\text{mem}}(t) = \sqrt{2(1-t)/t}$ memoryless	$\mathbb{R}_{\geq 0}$
Δt	Euler step size, $\Delta t = t_{i+1} - t_i$	$\mathbb{R}_{>0}$

1

Introduction

Text-to-image (T2I) generation has moved in a few years from a research novelty to a tool that millions use to turn a sentence into a picture, and the diffusion models behind it now render texture, lighting, and material with striking realism [11, 5]. What has not advanced at the same pace is control over the *composition* of a scene. Given a prompt that names several objects and dictates how they are arranged relative to one another, current models routinely get the arrangement wrong: they drop an object, merge two into one, or reverse the relation that was asked for. A request as plain as a dog positioned to the right of a teddy bear is satisfied less than half the time on Stable Diffusion 3.5, and the failure rate climbs as a scene grows more crowded and its relations more numerous.

This gap matters because the relations between objects - one beside another, one behind another, one resting on another - are much of what separates a coherent picture of a scene from an assembly of individually plausible parts. A generator that cannot be trusted to honor those relations cannot be steered, and so serves as a source of attractive but unpredictable images rather than as a controllable design instrument.

Part of the difficulty is architectural. The current generation of backbones, among them Stable Diffusion 3.5 [11] and FLUX [5], are multi-modal diffusion transformers (MM-DiT) in which a single joint-attention operator mixes image and text tokens in both directions at every layer. This design sharpens image quality, but it dissolves the one-way image-to-text cross-attention map that a substantial earlier line of training-free layout methods relied on to steer UNet diffusion models region by region [7, 35, 23, 34], so those techniques do not carry over.

Existing answers for MM-DiT divide along the familiar line between training and inference. Conditioning the backbone on layout inputs and fine-tuning it [38, 18, 37] is reliable but expensive and ties the result to one model, while training-free methods intervene during sampling instead, clamping the joint-attention matrix or stitching separately denoised per-box patches into the canvas [8, 2, 17, 39]. The latter assemble a scene region by region rather than steering all entities together, tend to leave seams at box boundaries, and are usually demonstrated on a single relation drawn from the four compass directions. A principled, training-free way to enforce many simultaneous relations among several entities on MM-DiT has been missing.

This thesis closes that gap with FOCAL, a training-free controller that treats layout guidance as a problem of stochastic optimal control (SOC) over the diffusion sampler. At each denoising step FOCAL nudges the model's predicted velocity by a single closed-form term, read directly from the network's own attention and derived from a cost that at once pulls each entity's attention into its assigned box and keeps the different entities' attention apart. The correction perturbs only the sampling velocity and leaves every weight frozen, so one controller serves several backbones unchanged and brings the SD3.5 and FLUX base models to a level competitive with far larger state-of-the-art systems [6, 33]. The formulation extends the memoryless SOC schedule of Adjoint Matching [10] and the attention-disentanglement controller of FOCUS [4] from a separation-only objective to a joint objective that also encodes spatial placement.

This thesis aims to design a principled, training-free controller that enforces arbitrary spatial relations among several entities at once and carries across MM-DiT backbones without retraining; and to evaluate that controller rigorously, including against a benchmark built specifically to isolate the spatial behavior a region-based controller can influence.

The remainder of this thesis follows the structure of a thesis by publication. Chapter 2 presents the scientific article, which sets out the motivation, method, and experiments of FOCAL in self-contained form. Chapter 3 develops in depth the background the article necessarily compresses, covering rectified-flow sampling on MM-DiT and the stochastic-optimal-control machinery the method rests on. Chapter 4 draws the work together, weighs its limitations, and sets out directions for future research.

2

Scientific Article

Training-Free Spatial Control for Multi-Entity Text-to-Image Generation

Vladimir Petkov[1] Thuy Nhat Vy Dinh[1] Athanasios Masouris[2] Sieger Falkena[2]
Tejas Gokhale[3] Hamid Palangi[†,4] Hadi Jamali-Rad[†,1,2]

[1]Delft University of Technology [2]Shell Information Technology International
[3]University of Maryland, Baltimore County [4]Google Research

† Equal co-advising

Abstract

Text-to-image models remain unreliable when asked to place two or more entities in a scene and respect the spatial relations between them. We present FOCAL (Flow Optimal Control for Alignment in Layout-Guided T2I Diffusion), a training-free, plug-and-play spatial guidance approach. Following a stochastic optimal control formulation, the controller applies a single-pass velocity correction at every denoising step, which is novel for spatial guidance. FOCAL introduces a unified running cost that couples attention disentanglement with layout conditioning in a single objective. To evaluate multi-entity placement, we curate SGE-MINI, a 1,000-prompt spatial benchmark. Since it acts only on the sampling velocity, the same controller transfers to multiple backbones with no retraining, substantially improving spatial alignment and bringing a base model to a level competitive with much larger state-of-the-art (SOTA) models.

1. Introduction

Modern text-to-image (T2I) models built on multi-modal diffusion transformers (MM-DiT), including Stable Diffusion 3.5 (SD3.5) [19] and FLUX.1-dev [8], remain unreliable on *spatial* constraints: which entity appears where, in what relative position, and how multiple entities are separated. Even a prompt as simple as `<a photo of a dog right of a teddy bear>` may produce the opposite positioning more often than not on SD3.5 (Figure 1). Such spatial understanding is crucial to realistically depict the human world and improve overall fidelity. Spatial-layout controllers for diffusion models traditionally fall into two families: training-based and training-free. *Training-based* methods condition a UNet or MM-DiT on additional structured adapters and require fine-tuning [21]. *Training-free* methods modify inference: by injecting per-region attention masks, by transplanting noisy patches from per-bbox sub-prompts [1, 28], or by optimising a hand-crafted loss on cross-attention statistics [6, 16, 39, 51, 53]. Most methods focus on single-relation settings and a limited scope of relations (usually the four relations: up, down, left, and right). We introduce FOCAL, a training-free, plug-and-play spatial-layout controller for MM-DiT (Figure 2). FOCAL rests on two ideas: it steers generation through a stochastic optimal control (SOC) correction applied to the sampler at inference, and it minimises an attention-based running cost that simultaneously separates entities and places each one inside its target bounding box to form a coherent layout. The method addresses arbitrary spatial relations in a complex multi-entity setting. Our contributions are three:

- **SOC for spatial layout on MM-DiT.** FOCAL is the first explicit SOC formulation of layout-based guidance on MM-DiT. Rather than running an inner optimisation at each step, we correct the sampler's velocity once per denoising step with a closed-form update derived from the combined disentanglement and layout objective.
- **Unified disentanglement and spatial running cost.** A single optimisable cost that enforces attention disentanglement and spatial bounding-box conditioning simultaneously, on the hypothesis that the two are interconnected: an entity can only be reliably placed inside its bounding box once its attention has been separated from those of other entities, and that separation must persist across the trajectory rather than being enforced at a single step.
- **SGE-MINI spatial benchmark.** A 1,000-prompt benchmark we curate from SPATIALGENEVAL [46], restricted to the five guidable spatial categories: Object, Position, Orientation, Proximity, and Occlusion. Prompts were shortened to isolate controllable spatial reasoning for fine-grained evaluation of multi-entity layout adherence in complex realistic settings.

On SGE-MINI and POSEVAL [1], FOCAL lifts the unmodified SD3.5 and FLUX.1-dev baselines' spatial accuracy by sig-

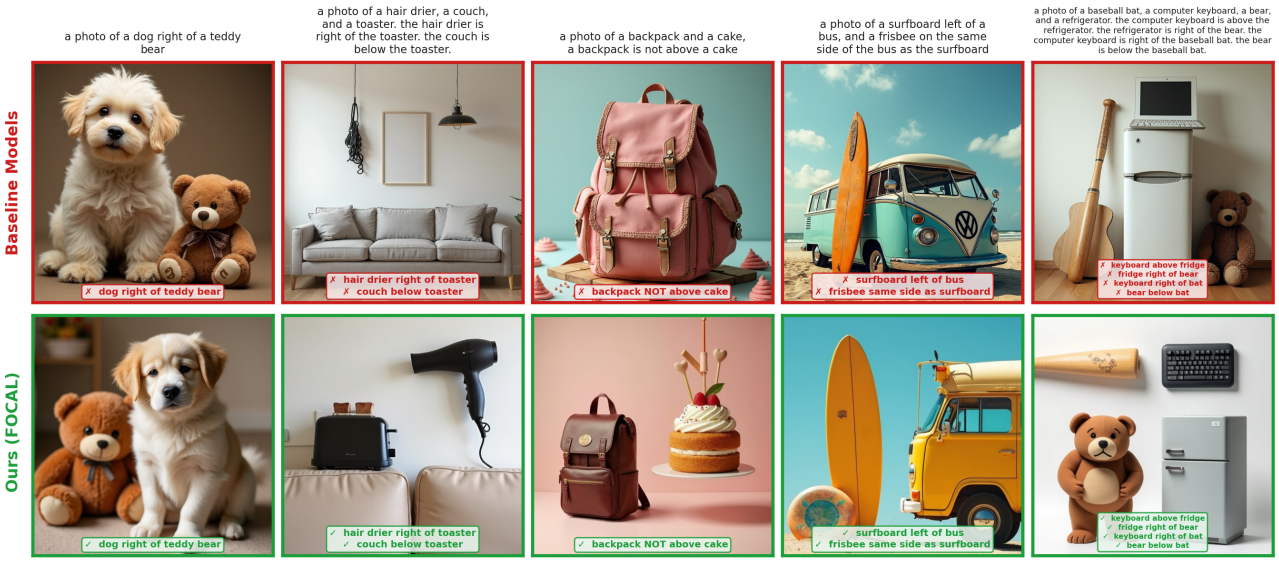


Figure 1. **FOCAL enforces spatial relations at inference, with no training.** Top row: unmodified Diffusion Transformer models often arrange entities incorrectly. Bottom row: applying FOCAL to the same prompts satisfies every relation. The prompts span single relations (*right of*, *below*), negation (*not above*), same-side constraints, and multi-entity prompts with several simultaneous relations; each satisfied (✓) or violated (✗) relation is listed below its image.

nificant margins, matching the performance of larger SOTA models without requiring any training.

2. Related Work

Controlling Diffusion Models. Conditioning a pretrained text-to-image model toward an objective happens at one of three stages. (i) During pretraining, the condition is learned into the weights: ControlNet [55] and GLIGEN [31] add trainable modules for edges, depth, or grounding boxes, requiring paired supervision and binding the controller to one backbone. (ii) After pretraining, the weights can be fine-tuned toward a reward by reinforcement learning [7], reward backpropagation [15], or preference tuning [44], all prone to reward overfitting. (iii) Training-free approaches instead steer the sampling trajectory with gradient additions toward an external objective [3, 17], or select from a reward-tilted posterior without differentiating it [30]; in both, the correction strength is heuristic. Adjoint Matching [18] shows the memoryless schedule fixes that strength while keeping the dynamics unbiased, but its objective is external to the model. OC-Flow [45] runs SOC over the whole trajectory toward an image-space reward, and FOCUS [6] applies SOC for attention disentanglement in multi-entity scenes with no spatial objective. FOCAL also runs SOC over an attention-based objective, but for the first time with a multi-entity spatial objective, unifying disentanglement and layout guidance.

Spatial Layout Guidance. Training-free U-Net layout methods perturb either the attention maps or the latent. Attention methods amplify or constrain each entity's map to its region [10, 56], penalise mass outside the box [39, 51, 53], or mask cross-subject attention to stop entities blending [16]. Latent methods denoise each region separately and blend the overlaps [4]. A further line retrains the model, for instance, editing text embeddings to fix a directional relation [50]. None transfer to MM-DiT, where a single joint-attention matrix couples every image and text token. On MM-DiT, regional masks [11, 14] and hard-mask injection [57] clamp the joint-attention matrix but introduce boundary artefacts, while GrounDiT [28] and Stitch [1] transplant a separately denoised patch per box through regional prompts, compositing rather than guiding. CreatiLayout [54] instead trains dedicated layout modules. FOCAL, rather than editing attention or the latent region by region, applies one training-free layout-based velocity correction that updates all entities jointly.

3. Preliminaries

3.1. Rectified-flow sampling on MM-DiT

A rectified-flow T2I model [19, 36, 37] synthesizes an image by integrating the ordinary differential equation (ODE):

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v_\theta(X_t, t, c), \qquad X_0 \sim \mathcal{N}(0, I), \tag{1}$$

over $t \in [0, 1]$ in a frozen VAE latent space, then decodes the terminal latent X_1 to image space, where c is a conditioning signal. The velocity field v_θ is trained by a conditional

flow matching (FM) objective with $(\alpha_t, \beta_t) = (t, 1-t)$, i.e. straight paths from noise to data: $X_t = (1-t)X_0 + tX_1$. At inference, (1) is integrated by Euler steps $X_{t_{i+1}} = X_{t_i} + (t_{i+1} - t_i)\, v_\theta(X_{t_i}, t_i, c)$ on $0 = t_0 < \cdots < t_N = 1$.

The velocity field is computed by an MM-DiT [8, 19, 38] that processes the concatenated sequence of image and text tokens through K *joint-attention* blocks. Each block attends jointly over both streams, producing one matrix over all $(N_{\text{img}} + N_{\text{txt}})$ tokens. Its image-to-text submatrix gives, for each text token, a distribution over image positions (a spatial map), the differentiable signal our running cost reads.

3.2. Stochastic Optimal Control for FM-SDE

Controlled Dynamics and Cost. Following [6, 18], a time-dependent control $u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ turns the ODE (1) into the controlled SDE $\mathrm{d}X_t^u = [\, b(X_t^u, t) + \sigma(t)\, u(X_t^u, t)\,]\,\mathrm{d}t + \sigma(t)\,\mathrm{d}B_t$, with initial noise $X_0^u \sim \mathcal{N}(0, I)$, B_t an $\mathbb{R}^d$-valued Brownian motion, $\sigma(t)$ a diffusion schedule, and b the FM base drift. We choose u to minimize a cost $J(u)$ with three terms: (i) a running cost $f(X_t^u, t)$ scoring the objective at each step (e.g. entity disentanglement [6]), (ii) a terminal cost g (e.g. a PickScore [27]) set to 0 following [6], and (iii) a control penalty $\frac{1}{2}\|u\|^2$ whose expected integral is the KL divergence from the base sampler [5]:

$$J(u) = \mathbb{E}_{X^u \sim P^u}\Big[\int_0^1 \big(\tfrac{1}{2}\|u\|^2 + f(X_t^u, t)\big)\,\mathrm{d}t + g(X_1^u)\Big]. \quad (2)$$

Optimal Control Computation. By Pontryagin's minimum principle [40], the optimal control minimizes the Hamiltonian $\mathcal{H}(x, u, a, t) = \frac{1}{2}\|u\|^2 + f(x,t) + a(t)^\top\big(b(x,t) + \sigma(t)u\big)$, the sum of the running cost $\frac{1}{2}\|u\|^2 + f(\cdot)$ and the inner product of the costate $a(t) \in \mathbb{R}^d$ with the controlled trajectory's drift $b(\cdot) + \sigma(\cdot)u$. The costate is the gradient of the value function $V(x,t)$, the smallest total cost still attainable from state x at time t to $t = 1$, so $a(t) = \nabla_x V$ is the sensitivity of the value to the state. Strict convexity in u yields the minimizer $u^* = -\sigma(t)\, a(t)$, and the costate satisfies the adjoint equation $\dot{a}(t) = -\nabla_x b(X_t^u, t)^\top a(t) - \nabla_x f(X_t^u, t)$, given $a(1) = \nabla_x g(X_1^u)$, integrated backward in time. This requires the full forward trajectory X_1^u, so following FOCUS [6] we drop $\nabla_x b^\top a$ (with $g \equiv 0$ giving $a(t) = \int_t^1 \nabla_x f(X_s^u, s)\,\mathrm{d}s$) and freeze $\nabla_x f$ over $[t, 1]$, so $a(t) \approx (1-t)\,\nabla_x f(X_t^u, t)$ and

$$u_t^* \approx -\sigma(t)(1-t)\,\nabla_x f(X_t^u, t). \quad (3)$$

Control to Velocity. The control u^* acts on the drift b of the SDE, but at inference we run the deterministic ODE, so we recast it as a velocity correction. An SDE and the ODE produce the same distribution of the state X_t at every t when the deterministic velocity equals the drift minus the score term $\frac{1}{2}\sigma^2(t)\,\nabla_x \log p_t$ (Fokker–Planck [41, 43]). For the uncontrolled SDE (drift b), this rule returns the model's own velocity v_θ, i.e. $b = v_\theta + \frac{1}{2}\sigma^2(t)\,\nabla_x \log p_t$. Applying it to the controlled drift $b + \sigma(t)u^*$ and using the score shift $\nabla_x \log p_t^u - \nabla_x \log p_t = u_t^*/\sigma(t)$ [5, 18] with (3)

$$v_t^u = v_\theta + \tfrac{1}{2}\sigma(t)\, u_t^* = v_\theta - \eta(t)\,\nabla_x f(X_t^u, t), \quad (4)$$

where $\eta(t) = \frac{1}{2}\sigma^2(t)(1-t)$.

The Memoryless Schedule. The optimal control reweights each generated image's likelihood by $e^{-\int_0^1 f(\cdot)\,\mathrm{d}t}$, making low-cost images more likely, so to sample images from $X_0 \sim \mathcal{N}(0, I)$ under the controlled dynamics the starting noise must stay Gaussian after reweighting. This holds only when X_0 and X_1 are independent in the base process, which a unique noise schedule enforces: $\sigma_{\text{mem}}^2(t) = 2\beta_t\big(\frac{\dot{\alpha}_t}{\alpha_t}\beta_t - \dot{\beta}_t\big)$ [18]. For rectified flow this is $2(1-t)/t$, so

$$\eta(t) = \tfrac{1}{2}\sigma_{\text{mem}}^2(t)(1-t) = \frac{(1-t)^2}{t}. \quad (5)$$

The schedule is strong early, when the sample is mostly noise, and weak later, once the content is set.

4. Method

4.1. From Prompts to Layouts

Establishing the correct layout is a crucial part of successful compositional generation [22]. Breaking this into scene-graph generation followed by layout synthesis leads to a more constraint-aligned outcome [20]. Notably, we use LLMs for both scene-graph and layout generation. Given a prompt c `<a teddy bear to the left of a baseball glove>` with $E = 2$ entities, two consecutive LLM calls produce the scene-graph triplets together with the per-entity token groups $\{\mathcal{T}_e\}_{e=1}^E$ (here `<teddy bear>` and `<baseball glove>`) and the associated bounding boxes $\{\mathcal{B}_e\}_{e=1}^E$.

Scene-Graph Parsing. The first stage maps c to a scene graph $\mathcal{G}(c) := (\mathcal{E}(c), \mathcal{R}(c))$ of entities $\mathcal{E}(c)$: (`<a teddy bear>` and `<baseball glove>`) and subject-predicate-object triplets $\mathcal{R}(c)$: `<teddy bear, left, baseball glove>`, following Kim et al. [26]. Notably, repeated entities from the prompt are disambiguated with numeric suffixes (`<bear_1>`, `<bear_2>`). The same call maps each entity to its token group $\mathcal{T}_e$: the parser additionally receives the prompt's exact parser token sequence as a numbered list and returns, per entity, the indices of all tokens that refer to it, including pronouns, synonyms, and paraphrases (`<the gardener>` ↔ `<she>`).

Layout Generation. The second stage follows Feng et al. [21]: an LLM call takes the prompt c and scene graph $\mathcal{G}(c)$ and generates a layout with normalized coordinates $\{(e, x_e, y_e, w_e, h_e)\}_{e \in \mathcal{E}(c)}$, with entity bounding boxes $\mathcal{B}_e = [x_e, x_e + w_e] \times [y_e, y_e + h_e] \subseteq [0,1]^2$. The

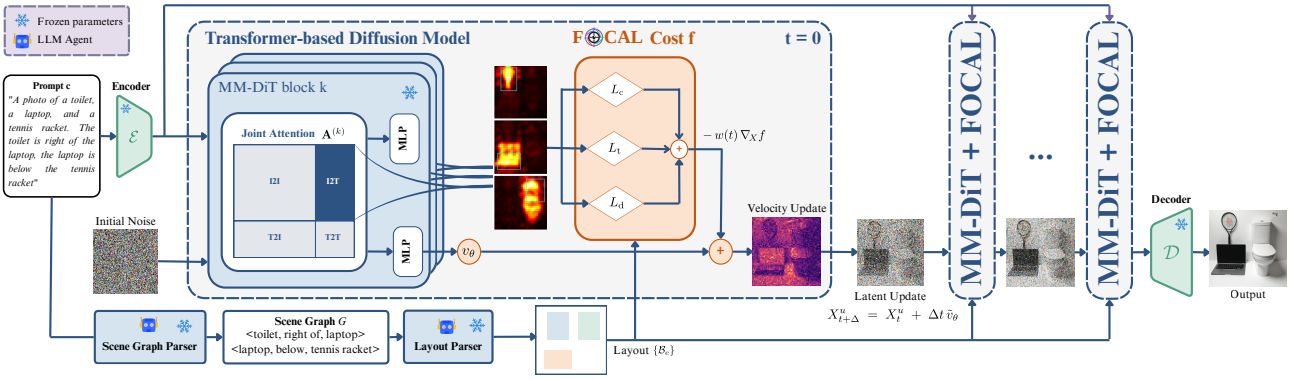


Figure 2. **The FOCAL pipeline.** An offline LLM stage (bottom, once per prompt) turns the prompt into entities and bounding boxes $\{\mathcal{B}_s\}$. During sampling, the frozen MM-DiT denoises the latent while we read its per-entity image-to-text attention A_e, score it with a running cost $f(\cdot)$, and correct the velocity as $\tilde{v}_\theta = v_\theta - w(t)\,\nabla_X f$ under a memoryless schedule. No weights are trained.

layout places every entity, satisfies all triplets, assigns plausible sizes, and avoids heavy overlap unless the relation implies it (<on>, <in front of>). Both stages share twenty expert-annotated in-context exemplars across six arrangement archetypes (horizontal, vertical, depth, center, proximity-surface, complex). The layout and token mappings are then fixed and held constant across all steps.

4.2. Attention Capture from MM-DiT

We read out the image-to-text submatrix of the joint-attention blocks [48]. On a forward pass with latent X_t^u, block k produces attention tensor $\mathbf{A}^{(k)}(X_t^u, t) \in \mathbb{R}^{N_{\text{img}} \times N_{\text{txt}}}$ whose (p, j) entry is the attention from image position p to text token j. We average across blocks,

$$\mathbf{A}(X_t^u, t) \;=\; \frac{1}{K}\sum_{k=1}^{K} \mathbf{A}^{(k)}(X_t^u, t), \tag{6}$$

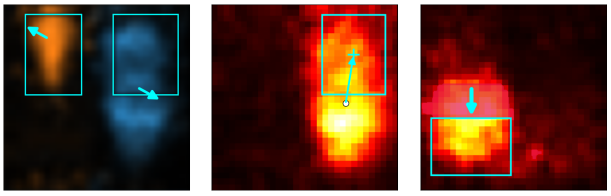
and apply a Gaussian smoothing with $\sigma = 1.5$ to suppress per-patch noise. For each text token j, the per-token spatial map $\mu_j(p) \propto \mathbf{A}(X_t^u, t)_{p,j}$, normalized so $\sum_p \mu_j(p) = 1$, is the distribution over image positions implied by attention to token j; averaging over the tokens of entity e gives the per-entity map $\bar{\mu}_e = \frac{1}{|\mathcal{T}_e|}\sum_{j\in\mathcal{T}_e} \mu_j$.

4.3. The Composite Running Cost

The running cost scores how well each $\bar{\mu}_e$ agrees with $\mathcal{B}_e$ and how well different entities' maps are disentangled. It combines a disentanglement term with two spatial terms, all computed from the readout $\mathbf{A}(X_t^u, t)$:

$$f(X_t^u, t) \;=\; w_{\text{d}}\, L_{\text{d}} \;+\; w_{\text{t}}\, L_{\text{t}} \;+\; w_{\text{c}}\, L_{\text{c}}, \tag{7}$$

with non-negative weights. $\mathcal{L}_{\text{d}}$ separates the entities' maps, $\mathcal{L}_{\text{t}}$ pulls each toward its box center, and $\mathcal{L}_{\text{c}}$ contains each within its box. The terms are complementary: disentanglement alone ignores placement, while the spatial terms alone let entities fuse and leak attributes. Figure 3 shows the effect of those terms.



$\mathcal{L}_{\text{d}}$ $\mathcal{L}_{\text{t}}$ $\mathcal{L}_{\text{c}}$

Figure 3. **Geometric action of the three spatial guidance terms**, overlaid on cross-attention maps. Cyan boxes mark the target layout; arrows indicate the gradient direction each term induces on the attention. $\mathcal{L}_{\text{d}}$ penalizes overlap between distinct entities' attention, repelling their centroids apart; $\mathcal{L}_{\text{t}}$ aligns the entity's attention centroid with its target box center; $\mathcal{L}_{\text{c}}$ penalizes attention mass outside the box, pulling it inward.

Attention Disentanglement. We adopt the FOCUS attention-binding disentanglement objective [6]. The per-token maps μ_j for the $T = \sum_e |\mathcal{T}_e|$ selected tokens form the rows of $\mathbf{Q} \in \mathbb{R}^{T\times N_{\text{img}}}$ ($\mathbf{Q}_j = \mu_j$). Following Lin [33], we use the Jensen-Shannon divergence (JSD), the mean KL of a set of distributions to their mean. The disentanglement loss is

$$L_{\text{d}} \;=\; \tfrac{1}{2}\big(\, L_{\text{coh}} \;+\; L_{\text{sep}}\,\big), \tag{8}$$

combining an intra-group coherence term $L_{\text{coh}} = \frac{1}{E}\sum_e \text{JSD}\big(\{\mathbf{Q}_j : j \in \mathcal{T}_e\}\big)$ (same-entity tokens should attend to the same region) and an inter-group separation term $L_{\text{sep}} = 1 - \text{JSD}(\{\bar{\mu}_1, \ldots, \bar{\mu}_E\})$ (different entities, different regions).

Centroid Translation. The translation term $L_{\text{t}} = \frac{1}{E}\sum_e \|\text{cent}(\bar{\mu}_e) - \mathbf{c}_e\|_2^2$, with $\text{cent}(\bar{\mu}_e) = \sum_p \bar{\mu}_e(p)\,\mathbf{p}$, pulls each entity's attention centroid to the center $\mathbf{c}_e$ of $\mathcal{B}_e$, ignoring spread.

Algorithm 1 FOCAL

Require: Prompt c; token groups $\{\mathcal{T}_e\}$ and boxes $\{\mathcal{B}_e\}$ (§4.1); weights $w_\mathrm{d}, w_\mathrm{t}, w_\mathrm{c}, \lambda$; steps $0 = t_0 < \cdots < t_N = 1$.

Sample $X^u_{t_0} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
for $i = 0, \ldots, N-1$ **do**
  $\mathbf{A}(X^u_{t_i}, t_i) \leftarrow \frac{1}{K}\sum_k \mathbf{A}^{(k)}(X^u_{t_i}, t_i)$ ▷ Eq. (6)
  $v_\theta(X^u_{t_i}, t_i, c) \leftarrow$ base velocity (with CFG)
  $L \leftarrow f(X^u_{t_i}, t_i)$ ▷ Eq. (7)
  $g \leftarrow \nabla_{X^u_{t_i}} L$ ▷ autograd through $\mathbf{A}$
  $\tilde{v}_\theta \leftarrow v_\theta - \lambda\,(1-t_i)^2/t_i \cdot g$ ▷ $w(t_i)\,g$
  $X^u_{t_{i+1}} \leftarrow X^u_{t_i} + (t_{i+1} - t_i)\,\tilde{v}_\theta$
end for
return $\mathcal{D}(X^u_{t_N})$ ▷ VAE decode

Distance-weighted Containment. The containment term penalizes mass outside the box in proportion to how far outside it lies. With positions $\mathbf{p} = (p_x, p_y)$ and the box in the same normalized $[0,1]^2$ frame, let $\delta(q; a, b) = \max(a - q,\, q - b,\, 0)$ be the distance of a coordinate q outside $[a, b]$, zero inside and growing linearly outside. The ℓ_∞ exterior distance to $\mathcal{B}_e = [x_0^{(e)}, x_1^{(e)}] \times [y_0^{(e)}, y_1^{(e)}]$ is then $d_e(p) = \max\big(\delta(p_x; x_0^{(e)}, x_1^{(e)}),\, \delta(p_y; y_0^{(e)}, y_1^{(e)})\big)$, and

$$L_\mathrm{c} \;=\; \frac{1}{E}\sum_{e=1}^{E}\Bigg(\sum_{p}\bar{\mu}_e(p)\,d_e(p)\Bigg)^2. \tag{9}$$

We use ℓ_∞ rather than ℓ_2 because a peak lies outside the box to the extent it violates *either* axis interval: ℓ_∞ takes the larger per-axis violation and penalizes a whole edge equally, whereas ℓ_2 over-penalizes diagonal departures.

4.4. Velocity Update

Substituting the memoryless schedule (5) and the composite cost (7) into (4) gives the per-step update

$$\tilde{v}_\theta(X^u_{t_i}, t_i, c) = v_\theta(X^u_{t_i}, t_i, c) - w(t_i)\,\nabla_{X^u_{t_i}} f(X^u_{t_i}, t_i), \tag{10}$$

where v_θ is the classifier-free-guidance base velocity at guidance strength s. The weight $w(t) = \lambda\,\eta(t) = \lambda\,(1-t)^2/t$ adds one scalar λ that sets the controller's global step size independently of the cost weights $(w_\mathrm{d}, w_\mathrm{t}, w_\mathrm{c})$. The corrected $\tilde{v}_\theta$ replaces v_θ in the Euler step, and the final latent $X^u_{t_N}$ is decoded by the VAE $\mathcal{D}$.

Architectural Portability. Algorithm 1 applies to any MM-DiT: it needs only a per-block image-to-text submatrix that can be averaged into (6) and back-propagated through.

5. Experiments

The SGE-MINI benchmark. SPATIALGENEVAL, introduced by Wang et al. [46], spans ten question types. We build SGE-MINI on its spatial taxonomy, keeping the four spatial categories whose answers depend on where and how attention mass lands (*Position*, *Orientation*, *Proximity*, *Occlusion*) together with *Object* presence and discarding the five other categories (*Attribute*, *Layout*, *Comparison*, *Motion*, *Causal*), which probe properties region-based guidance cannot affect. We sampled 1,000 of the 1,230 prompts at random and refined them in two stages: an LLM applied a first automated pass, after which two expert annotators reviewed every prompt and its question-answer pairs. They added a distinguishing modifier where two objects would otherwise share an attribute, restored missing sentence subjects, corrected grammar, punctuation, and plural agreement, and named repeated countable entities individually (`<a first child, a second child, and a third child>` rather than `<three children>`), so that each instance is a separately referable entity. Since the prompts and question-answer pairs differ from the originals, we evaluated every model in Table 1 under the protocol below. On POSEVAL, the relations involve at most four objects in canonical configurations, so any reasonable layout, whether benchmark-provided or generated, converges to the same arrangement, and the comparison isolates the guidance mechanism. SGE-MINI's complex multi-entity prompts admit no canonical layout, so imposing a single generated layout on every guidance method would conflate layout synthesis with guidance; we therefore report the training-free layout baselines on POSEVAL only.

Setup. We evaluate FOCAL on SD3.5-Medium (2.5B) and FLUX.1-dev (12B). Both stages of the layout pipeline (§4.1) use GPT-5.1 with the twenty in-context exemplars. The cost weights of Eq. (7) are shared across backbones at the values selected in §6: $w_\mathrm{d} = 0.20$, $w_\mathrm{t} = 2$, $w_\mathrm{c} = 10$. Only the guidance strength λ is set per backbone, 500 on FLUX.1-dev and 1500 on SD3.5-Medium, since each MM-DiT has a different architecture and so needs a different overall guidance scale; the cost weights transfer unchanged. The centroid is computed from a Gaussian-blurred attention map ($\sigma = 1.5$) restricted to mass above its 90th percentile. On SD3.5-Medium, the correction is applied at every transformer block; on FLUX.1-dev it is restricted to blocks 17-54, following the vital-layer analysis of Zhou et al. [57]. The earliest and latest FLUX blocks carry global image structure, while the middle blocks render per-entity details, so steering only the middle band guides the entities. For POSEVAL, each prompt is generated with 12 deterministic seeds, organized as three meta-seed runs of four images; for SGE-MINI, each prompt is generated once.

POSEVAL. We also evaluate on POSEVAL [1], six 100-prompt tasks that extend GenEval's *Position* test to harder compositions. *2-Obj*, *3-Obj*, and *4-Obj* chain relations among multiple entities (e.g. `<the oven is left of the fork, the fork is below the bird>`); *PAB* binds an attribute to each object; *Neg* uses negated relations; and *Rel* states a relation relative to a third object.

Table 1. SGE-MINI per-category VLM-VQA accuracy. Gray rows (ours) apply FOCAL to the SD3.5-Medium and FLUX.1-dev backbones. *Sp.* is the mean over *Position, Orientation, Proximity, Occlusion*; *Avg* over all five axes. Icons denote method type: ○ diffusion model; □ unified understanding-and-generation model; △ training-free layout guidance.

Method	Avg	Obj	Pos.	Ori.	Prox.	Occ.	Sp.
○ SD1.5 [42]	0.274	0.295	0.242	0.349	0.293	0.189	0.268
○ PixArt [12]	0.380	0.472	0.322	0.486	0.377	0.241	0.357
○ Playground [29]	0.416	0.572	0.325	0.481	0.430	0.237	0.377
○ SD3 [19]	0.605	0.866	0.607	0.598	0.606	0.349	0.540
○ Sana [52]	0.613	0.825	0.655	0.598	0.626	0.361	0.560
○ SD3.5 [19]	0.626	0.872	0.644	0.595	0.643	0.375	0.564
○ FLUX.1-dev [8]	0.638	0.851	0.672	0.620	0.675	0.371	0.585
○ HiDream-O1-Image [9]	0.749	0.956	0.795	0.728	0.781	0.483	0.697
○ Qwen-Image [49]	0.758	0.972	0.805	0.734	0.795	0.485	0.705
□ UniPic2-9B [47]	0.637	0.849	0.664	0.636	0.664	0.374	0.585
△ SD3.5 + FOCAL	0.695	0.910	0.784	0.603	0.740	0.440	0.642
△ FLUX + FOCAL	0.698	0.904	0.763	0.643	0.741	0.440	0.647

Metrics. On SGE-MINI, each question is posed five times to a Qwen2.5-VL-72B-Instruct judge [2] and counted correct only when at least four of the five answers match the ground truth; we report per-category accuracy. POSEVAL is scored under the GENEVAL 2 framework [25], which replaces GENEVAL's fixed Mask2Former detector, whose verdicts the authors report diverged from human ratings as image models improved, with a VLM judge queried separately on each prompt element. We report two metrics from a single InternVL3.5-8B judge. VQAScore [35] is holistic: it returns the judge's probability that the image matches the whole prompt. Soft-TIFA [25], extending TIFA [24], decomposes the prompt into objects, attributes, and relations, scores one question per atom, and aggregates by geometric mean, so a single failed relation lowers the score. Soft-TIFA is our primary measure of placement accuracy as VQAScore may sometimes not distinguish partial from complete success. We report VQAScore as a complementary, coarser measure of overall alignment. Every row in Table 2 is computed by us under this protocol.

SGE-MINI results. Table 1 reports per-category accuracy. On SD3.5-Medium, FOCAL improves the baseline by 0.069 on *Avg* and 0.078 on *Sp.*, with the largest gain on *Position* (+0.140), the category the translation and containment terms (L_t, L_c) act on most directly; *Orientation*, which planar guidance does not target, barely moves (+0.008). Applied to FLUX.1-dev without modification, it yields a comparable improvement (+0.060 *Avg*, +0.062 *Sp.*), again largest on *Position* (+0.091). Both guided backbones surpass every other diffusion and unified model in the table, apart from HiDream-O1-Image and Qwen-Image, larger SOTA base models that serve here as a quality ceiling. Figure 4 shows this qualitatively.

POSEVAL results. Table 2 reports per-task Soft-TIFA and VQAScore. On SD3.5-Medium, FOCAL raises average Soft-TIFA from 0.31 to 0.82 and average VQAScore from 0.54 to 0.88, with the largest gains on the multi-object tasks (4-Obj VQAScore 0.61 to 1.00). Applied to FLUX.1-dev, FOCAL yields a comparable increase, to 0.80 Soft-TIFA and 0.86 VQAScore. The margin over the baselines widens as objects and relations accumulate. Both guided backbones attain the two highest average Soft-TIFA scores among all compared methods, training-free and training-based alike, and SD3.5-Medium also leads on average VQAScore (0.88), ahead of the much larger state-of-the-art HiDream-O1-Image (0.87).

6. Hyperparameter Ablation

We select FOCAL's free parameters on SGE-MINI-VAL, a held-out validation split of 100 SPATIALGENEVAL (SGE) prompts. They are disjoint from the SGE-MINI evaluation set, so the choice of configuration never affects the prompts used for the main evaluation. All validation runs use FLUX.1-dev with images generated at 512×512, and we score each configuration with the SGE categories and aggregate over the full 100-prompt set.

Cost Weights. Fixing the parser, the layout, and the guidance strength (λ=500; ablated below), we grid-search the three FOCAL cost weights of Eq. 7 on SGE-MINI-VAL, sweeping the containment weight $w_c \in \{0.5, 2, 3, 4, 5, 7, 8, 9, 10, 15, 20, 30\}$, the translation weight $w_t \in \{0.5, 1, 2, 4\}$, and the disentanglement weight $w_d \in \{0.05, 0.20, 0.35\}$. Table 3 (top) reports a representative slice around the selected region of the configurations with the highest spatial accuracy scores on the validation set. The full table is in the supplement. The configuration w_c=10, w_t=2, w_d=0.20 attains the best average and spatial accuracy (0.724 and 0.685) and leads on Proximity (0.780) and Occlusion (0.590), the two depth-dependent spatial relations a planar method least directly controls. We adopt this configuration for all main results.

Guidance Strength. Holding the selected cost weights fixed, we vary λ over $\{300, 400, 500, 600, 700\}$ (Table 3, bottom). Average and spatial accuracy peak at λ=500 (0.724 and 0.685). We use λ=500 on FLUX.1-dev throughout; higher λ does not improve accuracy and degrades image fidelity on qualitative inspection. The same scalar is set per backbone on a small validation check, since each architecture needs a different overall guidance scale; on SD3.5-Medium, we use $\lambda = 1500$, with the cost weights transferred unchanged.

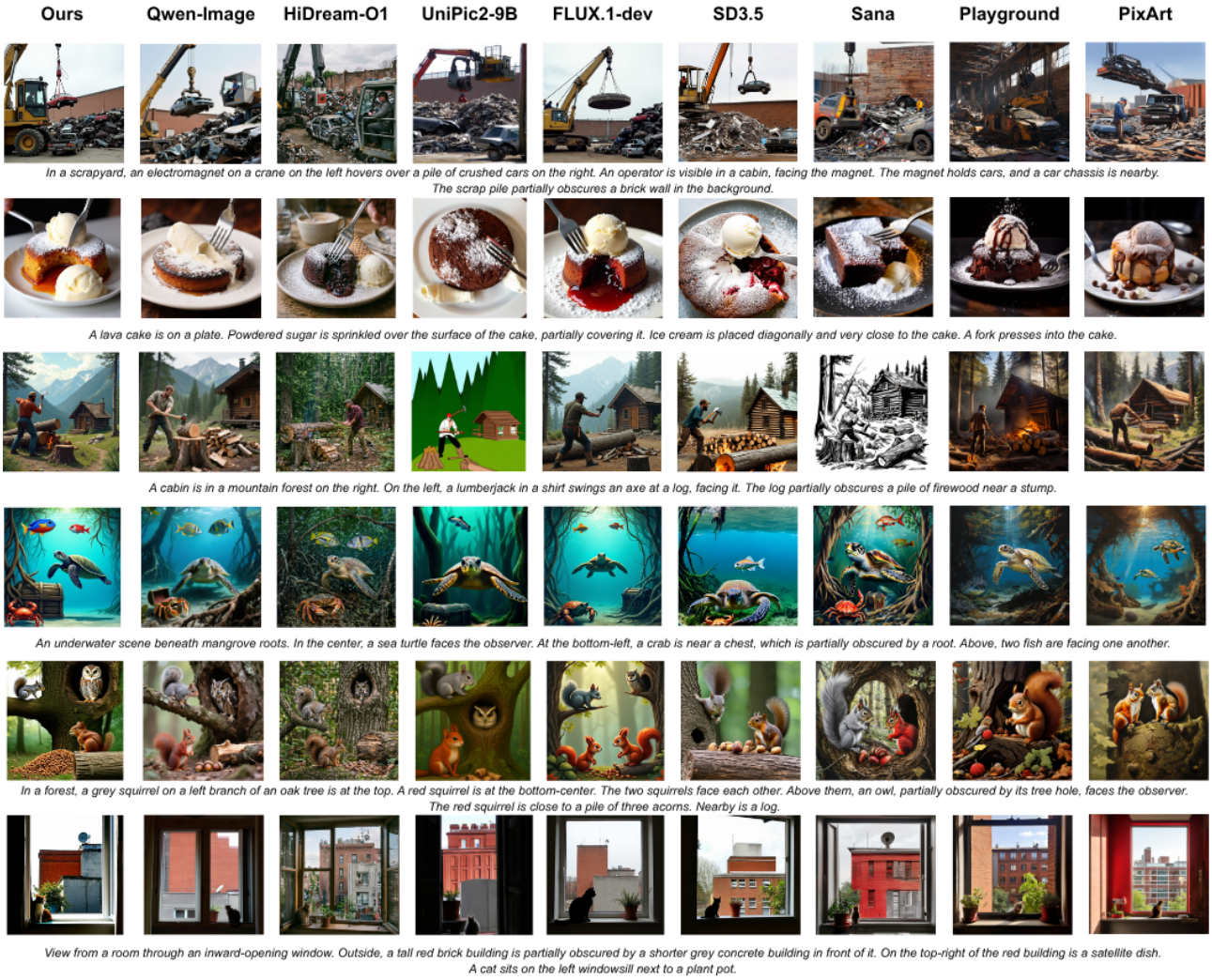


Figure 4. Qualitative comparison on multi-entity spatial-layout prompts from SGE-MINI. Each row is a prompt with multiple spatial relations; columns are FOCAL (ours) and the baselines Qwen-Image, HiDream-O1, UniPic2-9B, FLUX.1-dev, SD3.5, Sana, Playground, PixArt. FOCAL places every named entity in the requested configuration while preserving object identity and overall image quality.

Scene-Graph Parser. Table 4 compares five parsers (LLMs, fine-tuned scene-graph models, and a rule-based extractor) by spatial-relation recall on DiscoSG-DS over 18 relation groups we define (see supplement). We score recall because a missing spatial relation leaves an entity pair unguided and placed in the default manner, whereas an additional non-spatial relation is not central to the spatial scene layout. Few-shot GPT-5.1 leads on both micro and macro recall, with its largest margin on micro. The fine-tuned DiscoSG Refiner is competitive on macro recall but trails on micro, while the remaining fine-tuned and rule-based parsers under-generate sharply. We adopt GPT-5.1, and for consistency use the same model as the layout-planner agent that predicts bounding boxes from the parsed graph. The per-relation breakdown is in the supplement.

7. Discussion

FOCAL shows that a training-free correction to the sampling velocity is enough to enforce multi-entity spatial layout on a frozen MM-DiT. The disentanglement and placement are optimised as one running cost over the whole trajectory rather than as separate objectives. In a crowded scene, an entity cannot be moved cleanly into its box while its attention still overlaps with the others, so separating the entities and positioning them are not sequential steps but a single coupled problem; folding both into one cost lets the controller trade them off at every step.

Another finding is that guiding two-dimensional attention

Table 2. PosEval per task (InternVL3.5-8B judge, n=100/task). Gray rows are ours. Best per column **bold**, second underlined. Icons denote method type: ○ diffusion model; □ unified understanding-and-generation model; △ training-free layout guidance; ▲ training-based layout guidance.

	Soft-TIFA							VQAScore						
Method	2-Obj	3-Obj	4-Obj	PAB	Neg	Rel	Avg	2-Obj	3-Obj	4-Obj	PAB	Neg	Rel	Avg
○ SD1.5 [42]	0.19	0.04	0.00	0.04	0.33	0.03	0.11	0.23	0.17	0.16	0.12	0.36	0.10	0.19
○ Playground [29]	0.32	0.10	0.02	0.17	0.53	0.07	0.20	0.40	0.32	0.27	0.32	0.45	0.37	0.36
○ FLUX vanilla [8]	0.34	0.25	0.21	0.38	0.51	0.16	0.31	0.47	0.49	0.61	0.54	0.52	0.59	0.54
○ SD3.5 vanilla [19]	0.44	0.22	0.07	0.44	0.52	0.17	0.31	0.49	0.39	0.61	0.67	0.54	0.51	0.54
○ HiDream-O1-Image [9]	**0.98**	0.56	0.32	**0.94**	0.40	0.39	0.60	**1.00**	0.82	0.91	**0.99**	0.55	**0.95**	0.87
□ UniPic2-9B [47]	0.93	0.39	0.28	0.84	0.44	0.36	0.54	0.96	0.59	0.74	0.95	0.55	0.86	0.78
△ RAG-Diffusion [14]	0.68	0.29	0.28	0.26	0.71	0.36	0.43	0.70	0.50	0.58	0.47	0.50	0.44	0.53
△ RP-FLUX [11]	0.54	0.38	0.44	0.31	0.66	0.56	0.48	0.62	0.67	0.82	0.56	0.62	0.75	0.67
▲ CreatiLayout [54]	0.75	0.60	0.58	0.71	0.88	0.60	0.69	0.81	0.81	0.94	0.91	0.59	0.78	0.81
△ SD3.5 + FOCAL	0.92	0.71	0.82	0.73	**0.94**	**0.78**	**0.82**	0.94	0.85	**1.00**	0.90	**0.65**	0.93	**0.88**
△ FLUX + FOCAL	0.89	**0.73**	**0.84**	0.74	0.92	0.69	0.80	0.95	**0.91**	0.99	0.93	0.50	0.89	0.86

Table 3. Ablation on FOCAL cost weights (top) and λ (bottom). Best per column **bold** within each section. Gray rows are the chosen configuration.

Configuration				Summary		Per task				
w_c	w_t	w_d	λ	Avg.	Spatial	Obj.	Pos.	Ori.	Prox.	Occ.
5	1	0.05		0.652	0.618	0.790	0.730	0.630	0.680	0.430
5	1	0.20		0.674	0.637	0.820	0.740	**0.640**	0.730	0.440
5	1	0.35		0.684	0.645	0.840	0.760	0.630	0.710	0.480
5	2	0.05		0.666	0.625	0.830	0.730	0.620	0.670	0.480
5	2	0.20		0.662	0.627	0.800	0.740	0.600	0.730	0.440
5	2	0.35	500	0.664	0.625	0.820	**0.780**	0.600	0.700	0.420
10	1	0.05		0.686	0.640	0.870	0.750	0.600	0.700	0.510
10	1	0.20		0.676	0.635	0.840	0.740	0.610	0.710	0.480
10	1	0.35		0.692	0.647	0.870	0.720	0.630	0.690	0.550
10	2	0.05		0.690	0.652	0.840	0.770	0.610	0.710	0.520
10	2	0.20		**0.724**	**0.685**	**0.880**	0.760	0.610	**0.780**	**0.590**
10	2	0.35		0.672	0.632	0.830	0.730	0.590	0.710	0.500
			300	0.664	0.623	0.830	0.710	0.610	0.690	0.480
			400	0.684	0.645	0.840	0.700	0.600	0.740	0.540
10	2	0.20	500	**0.724**	**0.685**	**0.880**	**0.760**	0.610	**0.780**	**0.590**
			600	0.660	0.630	0.780	0.740	0.590	0.700	0.490
			700	0.694	0.657	0.840	0.750	**0.630**	0.700	0.550

Table 4. Scene-graph parser comparison: spatial-relation recall on DiscoSG-DS (100 examples, 18 relation groups). Gray row is the parser FOCAL adopts; best per column **bold**, second underlined.

Parser	Type	Micro	Macro
GPT-5.1	LLM	**0.682**	**0.550**
DeepSeek-V3.2	LLM	0.628	0.485
DiscoSG Refiner [34]	fine-tuned	0.583	0.546
FACTUAL-T5 [32]	fine-tuned	0.171	0.082
spaCy [23]	rule-based	0.109	0.109

may also, to an extent, improve inherently three-dimensional relations. The controller only moves where attention mass lands on the image plane, yet *Proximity* and *Occlusion*, which depend on depth ordering, improve substantially (+0.097 and +0.065 on SD3.5-Medium). Latent diffusion models are known to encode depth and figure-ground structure in their internal activations even when trained only on images [13]; placing the entities in the correct planar configuration appears to be enough to trigger this latent geometry, so the frozen model renders the occlusion ordering itself once the layout removes the ambiguity.

8. Conclusion

We presented FOCAL, a training-free controller that enforces bounding-box spatial layout on multi-modal diffusion transformers by correcting the sampling velocity under a stochastic-optimal-control formulation. The correction minimizes a single running cost that couples attention disentanglement with placement. We release SGE-mini, a 1,000-prompt realistic multi-entity spatial benchmark curated from SpatialGenEval. Because the controller acts only on the velocity, it transfers to SD3.5-Medium and FLUX.1-dev with no retraining and no change to the algorithm. It improves both base models on every category of SGE-mini and across all tasks of PosEval, where FOCAL attains the highest average Soft-TIFA placement accuracy of any compared method, ahead of far larger models such as the fully trained HiDream-O1-Image while requiring no training of its own.

References

[1] Jessica Bader, Mateusz Pach, María A. Bravo, Serge Belongie, and Zeynep Akata. Stitch: Training-free position control in multimodal diffusion transformers, 2025. 1, 2, 5

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. 6

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and

Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. 2

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 2

[5] Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2211.01364. 3

[6] Eric Bill, Enis Simsar, and Thomas Hofmann. FOCUS: Optimal control for multi-entity world modeling in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 1, 2, 3, 4

[7] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[8] Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024. 1, 3, 6, 8

[9] Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer, 2026. 6, 8

[10] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 2

[11] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024. 2, 8

[12] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. 6

[13] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*, 2024. arXiv:2306.05720. 8

[14] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 8

[15] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[16] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[18] Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.08861. 2, 3

[19] 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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1, 2, 3, 6, 8

[20] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Björn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 88–98, 2023. 3

[21] Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3

[22] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7986–7994, 2018. 3

[23] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in python. Zenodo, 2020. 8

[24] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6

[25] Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation, 2025. 6

[26] Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. LLM4SGG: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[27] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset

of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[28] Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. GrounDiT: Grounding diffusion transformers via noisy patch transplantation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[29] 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. arXiv:2402.17245. 6, 8

[30] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[31] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[32] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6377–6390, 2023. 8

[33] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1): 145–151, 1991. 4

[34] Shaoqing Lin, Chong Teng, Fei Li, Donghong Ji, Lizhen Qu, and Zhuang Li. DiscoSG: Towards discourse-level text scene graph parsing through iterative graph refinement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7837–7862, 2025. 8

[35] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 6

[36] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[37] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[39] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[40] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, 1962. 3

[41] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, 2 edition, 1996. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 8

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[44] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[45] Luran Wang, Chaoran Cheng, Yizhen Liao, Yanru Qu, and Ge Liu. Training free guided flow-matching with optimal control. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.18070. 2

[46] Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. In *International Conference on Learning Representations (ICLR)*, 2026. 1, 5

[47] Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, Chuanxin Tang, Zidong Wang, Yichen Wei, Liang Hu, Boyi Jiang, William Li, Ying He, Yang Liu, Xuchen Song, Eric Li, and Yahui Zhou. Skywork UniPic 2.0: Building kontext model with online rl for unified multimodal model, 2025. 6, 8

[48] Tianyi Wei, Dongdong Chen, Yifan Zhou, and Xingang Pan. Enhancing MMDiT-based text-to-image models for similar subject generation. *arXiv preprint arXiv:2411.18301*, 2024. 4

[49] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report, 2025. 6

[50] Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2403.20249. 2

[51] Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. R&B: Region and boundary aware zero-shot

grounded text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[52] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025. 6

[53] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[54] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. CreatiLayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 8

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[56] Peiang Zhao, Han Li, Ruiyang Jin, and S. Kevin Zhou. LoCo: Locally constrained training-free layout-to-image synthesis. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 9481–9490, 2025. 2

[57] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. DreamRenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 5

Training-Free Spatial Control for Multi-Entity Text-to-Image Generation

Supplementary Material

A. Additional qualitative comparisons

Figure 5. Qualitative comparison on multi-entity spatial-layout prompts from PosEval. Each row is a prompt with one or more spatial relations; columns are FOCAL on FLUX.1-dev (ours) and the baselines CreatiLayout, HiDream-O1, RP-FLUX, RAG-Diffusion, Playground, SD 1.5, and UniPic-2. FOCAL places every named entity in the requested configuration while preserving object identity and overall image quality.

B. LLM prompts and in-context exemplars

FOCAL turns a prompt into entities, per-entity token groups, and bounding boxes with two LLM calls (scene-graph parsing, then layout planning), both using GPT-5.1 with the twenty expert-annotated in-context exemplars in Appendix B.3. We show the **dual-tokenizer** variant: the scene-graph call is grounded against both the T5 token sequence (FLUX / SD3.5 backbone) and the Qwen2.5-VL post-drop token sequence (Qwen-Image backbone), returning one alignment map per tokenizer; for a single-backbone run only the relevant token list is supplied and the other alignment map is left empty.

B.1. Scene-graph parser — system prompt

```
You are a scene graph parser for text-to-image
generation. Your task is to extract a structured scene
 graph that describes positional relationships between
 entities in a descriptive text prompt. Additionally,
you are tasked to provide a token index mapping.

A scene graph consists of:
1. **Entities**: The objects, characters, and scene
elements mentioned in the text.
2. **Triplets**: Relationships between entities as [
subject, predicate, object] tuples.

Your focus is on SPATIAL relationships where things
are relative to each other. Also capture action
relationships (holding, wearing, riding) and surface/
contact relationships (on, sitting on, hanging from)
when they imply spatial positioning.

## Inputs

- A descriptive text prompt describing an image to be
generated, with relational and spatial information
encoded in text.
- A mapping for each word of the prompt to
corresponding token indices of a text encoder.



## Strict Output Format
Respond with ONLY a JSON object. Format:
{
  "entities": ["entity1", "entity2", ...],
  "token_alignment": {<entity>: <list of corresponding
 integer token indices>, ..},
  "token_alignment_qwen": {<entity>: <list of
corresponding integer token indices>, ..},
  "triplets": [
    ["subject", "predicate", "object"],
    ...
  ]
}

The two alignment maps use DIFFERENT, INDEPENDENT
index spaces: `token_alignment` indices come ONLY from
 the "T5 token indices" list, and `
token_alignment_qwen` indices come ONLY from the "Qwen
 token indices" list. Never mix them, and never emit
an index that does not appear in the corresponding
list. If a token list is not provided in the query,
output an empty object for its map.

You MUST never deviate from the output format. You are
 only allowed to output triplets as 3 items
corresponding to "subject" entity, "predicate"
relationship, and "object" entity.


## Rules

1. Extract ALL spatial relationships mentioned or
clearly implied in the text.
2. Use lowercase entity names. Use simple noun phrases
 matching the text.
3. Scene-level positions like "X is on the left" or "
On the left, X..." MUST produce a spatial triplet. Use
 another entity as the anchor:
   - If Y is mentioned on the opposite side -> ["X", "
left of", "Y"]
   - If Y is in the center -> ["X", "left of", "Y"]
   - If only X has a position -> ["X", "left of", <
most prominent other entity>]
4. Every entity with a stated position must appear in
at least one spatial triplet.
5. If A is left of B AND the text also says B is to
the right, include BOTH ["A", "left of", "B"] and ["B
", "right of", "A"].
6. For "between" relations, create TWO triplets: ["X",
 "between", "A"] and ["X", "between", "B"] where A and
 B are the flanking entities.
7. Disambiguate repeated entities with suffixes: "
person_1", "person_2".
8. Never produce contradictory predicates for the same
 entity pair.
9. Prefer specific predicates over generic ones: use "
sitting on" instead of "on" when the text says "
sitting on".
10. Use canonical forms (not synonyms):
    - "to the left of" / "on the left side of" -> "
left of"
    - "to the right of" / "on the right side of" -> "
right of"
    - "next to" / "adjacent to" -> "beside"
    - "close to" -> "near"
    - "over" (when meaning above) -> "above"
    - "in back of" -> "behind"
11. Avoid considering attributes/adjectives as part of
 the entity (e.g., "red car" -> "car", "distant tree"
-> "tree"). EXCEPTION: Unless those attributes/
adjective explicitly distinguish specific parts of the
 entity (e.g., "woman's leg", "tree's branch", "the
door of the car", "car front").
12. Every entity name MUST be an exact substring of
the prompt (case-insensitive). Never invent entity
names not present in the text.
13. Do NOT extract these as entities if they serve as
scene-level descriptors or viewpoint indicators and
hold no spatial meaning: observer, viewer, audience,
foreground, background, center, left side, right side,
 ground, floor, ceiling, surface, horizon, air, water,
 shadow, light, fog, reflection, sky, space, area,
edge.
14. Output ONLY a JSON object -- no explanation, no
markdown fences.

### Token Alignment
The query may include ONE OR TWO numbered token lists
("T5 token indices" and/or "Qwen token indices").
These are two different tokenizations of the SAME
prompt, so they use different index spaces. Produce a
SEPARATE alignment map for each list that is provided:
 indices from the "T5 token indices" list go into `
token_alignment`, and indices from the "Qwen token
indices" list go into `token_alignment_qwen`. The set
of entities and their references is identical across
both maps -- only the integer indices differ.

For each entity you identify, and for each provided
token list:
1. Locate every reference of it in the input prompt -
references can be implicit (e.g., pronouns (hers, they
), synonyms (small bear -> cub), identities (gardener
<-> human/woman/man))
2. Collect the token indices of those references FROM
THAT LIST
3. Add them as token alignment mappings for the
corresponding map
```

```
IMPORTANT:
- Only use indices that actually appear in the
respective numbered list; never emit an out-of-range
or cross-list index.
- Ensure that the corresponding token indices of all
references of an identified entity in the input prompt
 are included in the token alignment mapping.
- In case of prefix token overlap (e.g., "door" and "
doorman") ensure that entities are mapped to the
correct tokens based on sentence context.
- If we have multiple entities of same "type" (as in "
bird_1", "bird_2"), make sure that every reference of
the base type ("bird") in the prompt, and their
corresponding tokens, are mapped to the correct
identified entities.
- If there are descriptive attributes/adjectives (e.g
., "distant flag", "big tree", "red car"), add their
token indices in the token alignment mapping for the
corresponding entities ("flag", "tree", "car",
respectively)
- Paraphrasing MUST be recognized. For example, if in
a prompt you have "the bonsai tree's branch" and "the
branch of the bonsai tree", and they both contextually
 refer to the same entity "bonsai tree's branch", you
should map all of those token indices to that entity
in the token alignment mapping.
```

User message (per query). The prompt is followed by one or both numbered token lists; the parser returns the indices, per entity, that refer to it in each list:

```
Parse this prompt into a scene graph:

"<PROMPT>"

T5 token indices:
[0]_At [1]_ [2]a [3]_Railway [4]_Station ...      #
SentencePiece marker U+2581 rendered as '_'

Qwen token indices:
[0]A [1]_cat [2]_sitting ...                      #
byte-BPE space U+0120 -> '_', newline U+010A -> '<nl>'
```

In few-shot mode the twenty exemplars (A.3) are prepended as user/assistant turns before this query. Token-alignment is taught by the instructions only — the exemplars themselves carry no alignment.

B.2. Layout planner — system prompt

```
You are a spatial layout planner for text-to-image
generation. Given a scene description and its scene
graph (entities + spatial relationships), produce a
bounding box layout that positions each entity
according to the described spatial arrangement.

## Output Format

Respond with ONLY a JSON object. Format:
{
  "layout": [
    {"entity": "entity_name", "x": float, "y": float,
"w": float, "h": float},
    ...
  ]
}

Where x, y are the TOP-LEFT corner coordinates and w,
h are width and height, all normalized to [0, 1] (
image coordinates: x=0 is left, y=0 is top).

## Rules

1. Every entity from the scene graph must appear in
the layout and do not change the names of those
entities.
2. Respect ALL spatial relationships: "left of" means
lower x, "above" means lower y, "below" means higher y
, etc.
3. Entities should not overlap excessively unless the
relationship implies it (e.g., "on", "in front of", "
partially obscures", "obscured by").
4. Use reasonable sizes: main subjects 0.2-0.4, small
objects 0.1-0.2, backgrounds/scenes 0.5-0.8.
5. Keep all boxes within [0, 1] bounds.
6. Output ONLY JSON -- no explanation, no markdown
fences.
```

User message (per query):

```
Generate a layout for this scene:

Prompt: "<PROMPT>"
Entities: ["entity1", "entity2", ...]
Relationships: [["subject", "predicate", "object"],
...]
```

B.3. The twenty in-context exemplars

Both stages share the same twenty expert-annotated prompts, spanning six arrangement archetypes (Horizontal, Vertical, Depth, Centre, Proximity-surface, Complex). The scene-graph stage uses the entities + triplets (Appendix B.3.1); the layout stage additionally uses the bounding boxes (Appendix B.3.2).

B.3.1. Scene-graph exemplars (entities + triplets)

Exemplar 1 — Horizontal. *Prompt:* `<A serene Japanese garden unfolds in the soft golden afternoon light. On the left, a weathered stone lantern rises from a bed of emerald moss, its surface patched with lichen. To the right, a graceful red maple tree arches over a shallow koi pond, its crimson leaves drifting gently onto the still water. A narrow gravel path winds between the lantern and the tree, bordered by neatly trimmed boxwood hedges.>`

- Entities: `<stone lantern>`, `<moss>`, `<maple tree>`, `<koi pond>`, `<leaves>`, `<gravel path>`, `<hedges>`
- Triplets: `<stone lantern, left of, maple tree>`; `<maple tree, right of, stone lantern>`; `<stone lantern, on, moss>`; `<maple tree, above, koi pond>`; `<leaves, on, koi pond>`; `<gravel path, between, stone lantern>`; `<gravel path, between, maple tree>`; `<boxwood hedges, beside, gravel path>`

Exemplar 2 — Horizontal. *Prompt:* `<In a dimly lit medieval tavern, a grizzled dwarf sits on the left at a heavy oak table, clutching a frothy mug of ale. To the right, a cloaked elf leans against the stone wall, studying a tattered parchment map by candlelight. A roaring fireplace occupies the center`

of the back wall, casting dancing shadows across the rough wooden floor and warming the stale, smoky air that hangs between the weary travelers.>

- Entities: <dwarf>, <oak table>, <mug>, <elf>, <stone wall>, <back wall>, <parchment map>, <candle>, <fireplace>, <floor>, <smoky air>
- Triplets: <dwarf, left of, elf>; <elf, right of, dwarf>; <dwarf, sitting on, oak table>; <mug, on, oak table>; <dwarf, holding, mug>; <elf, leaning against, stone wall>; <elf, holding, parchment map>; <candle, beside, elf>; <fireplace, between, dwarf>; <fireplace, between, elf>; <fireplace, on, floor>; <candle, on, oak table>; <fireplace, in the center of, back wall>; <smoky air, between, dwarf>; <smoky air, between, elf>

Exemplar 3 — Horizontal. *Prompt:* <A wide desert highway stretches toward distant purple mountains under an endless turquoise sky. On the left side of the road, a rusted gas station sign leans at an angle, its faded letters barely legible. To the right, a lone cactus stands tall beside a crumbling adobe wall. A dusty red pickup truck is parked between them on the gravel shoulder, its bed loaded with wooden crates and coils of barbed wire.>

- Entities: <highway>, <mountains>, <sky>, <gas station sign>, <cactus>, <adobe wall>, <pickup truck>, <crates>, <barbed wire>, <gravel shoulder>
- Triplets: <mountains, under, sky>; <gas station sign, left of, highway>; <cactus, right of, highway>; <cactus, beside, adobe wall>; <pickup truck, between, gas station sign>; <pickup truck, between, cactus>; <pickup truck, parked on, gravel shoulder>; <gravel shoulder, beside, highway>; <gravel shoulder, between, adobe wall>; <gravel shoulder, between, cactus>; <crates, on, pickup truck>; <barbed wire, on, pickup truck>; <mountains, behind, highway>

Exemplar 4 — Horizontal. *Prompt:* <A vibrant farmer's market fills a cobblestone plaza on a sunny Saturday morning. On the left, a woman in a straw hat arranges baskets of ripe strawberries on a wooden stand draped with checkered cloth. To the right, an old man with a white beard sells jars of golden honey from a small folding table. Children scamper between the stalls while a busker plays guitar near the stone fountain at the far end.>

- Entities: <plaza>, <market>, <woman>, <strawberries>, <straw hat>, <wooden stand>, <checkered cloth>, <old man>, <honey>, <folding table>, <children>, <busker>, <guitar>, <fountain>, <stalls>
- Triplets: <market, on, plaza>; <woman, left of, old man>; <old man, right of, woman>; <woman, wearing, straw hat>; <woman, arranging, strawberries>; <strawberries, on, wooden stand>; <woman, next to, wooden stand>; <checkered cloth, on, wooden stand>; <old man, right of, wooden stand>; <honey, on, folding table>; <old man, next to, folding table>; <children, between, stalls>; <busker, near, fountain>; <busker, holding, guitar>

Exemplar 5 — Vertical. *Prompt:* <Inside a grand cathedral, enormous stained glass windows tower above the polished marble floor, casting kaleidoscopic patterns of vivid colored light across the wooden pews. A heavy iron chandelier hangs from the vaulted ceiling far overhead, its dozens of candles flickering softly. Below, a priest in white robes stands behind a carved wooden altar adorned with fresh lilies and a golden cross. Pigeons roost on the ledge beneath the rose window.>

- Entities: <stained glass windows>, <patterns of vivid colored light>, <wooden pews>, <marble floor>, <chandelier>, <ceiling>, <candles>, <priest>, <robes>, <altar>, <lilies>, <golden cross>, <pigeons>, <rose window>
- Triplets: <stained glass windows, above, marble floor>; <stained glass windows, casting, patterns of vivid colored light>; <patterns of vivid colored light, on, wooden pews>; <chandelier, hanging from, ceiling>; <candles, on, chandelier>; <chandelier, above, altar>; <priest, below, chandelier>; <priest, standing behind, altar>;

<priest, wearing, robes>; <lilies, on, altar>; <golden cross, on, altar>; <pigeons, below, rose window>; <altar, below, chandelier>

Exemplar 6 — Vertical. *Prompt:* <A towering ancient oak tree dominates a misty autumn hillside dotted with wildflowers. High above, a red-tailed hawk circles lazily against a pale grey sky streaked with fading sunlight. A wooden treehouse sits nestled among the thick middle branches, its rope ladder dangling below toward the leaf-strewn ground. Underneath the tree, a sleeping fox curls up on a pile of golden leaves beside a moss-covered boulder, sheltered from the cool breeze.>

- Entities: <oak tree>, <hillside>, <wildflowers>, <hawk>, <sky>, <treehouse>, <branches>, <rope ladder>, <ground>, <fox>, <leaves>, <boulder>
- Triplets: <oak tree, on, hillside>; <wildflowers, on, hillside>; <hawk, above, oak tree>; <hawk, below, sky>; <treehouse, on, branches>; <rope ladder, hanging from, treehouse>; <rope ladder, above, ground>; <fox, under, oak tree>; <fox, lying on, leaves>; <leaves, on, ground>; <boulder, beside, fox>; <treehouse, above, fox>

Exemplar 7 — Vertical. *Prompt:* <A massive waterfall thunders down a sheer cliff face into a churning pool of turquoise water far below. Lush tropical ferns cling to the wet rocks on the side of the cliff, their fronds dripping with mist. Above the falls, a rickety wooden bridge spans the river, and a lone hiker stands on it, gazing down at the spectacle. Rainbow mist rises from the base of the waterfall, catching the midday sun.>

- Entities: <waterfall>, <cliff>, <pool>, <ferns>, <rocks>, <bridge>, <river>, <hiker>, <mist>
- Triplets: <waterfall, on the side of, cliff>; <pool, below, waterfall>; <ferns, on the side of, cliff>; <ferns, on, rocks>; <bridge, above, waterfall>; <bridge, across, river>; <waterfall, in, river>; <hiker, standing on, bridge>; <hiker, above, pool>; <hiker, above, waterfall>; <hiker, above, river>; <mist, above, pool>; <bridge, above, river>

Exemplar 8 — Depth. *Prompt:* <A narrow cobblestone alley winds through an old Italian village at dusk. In the foreground, a black cat perches on a crumbling brick wall, its green eyes glowing in the fading light. Behind it, a woman in a blue dress hangs laundry from a line stretched between two ochre-colored buildings. Far in the background, the bell tower of a church rises against a darkening sky streaked with bands of amber and violet.>

- Entities: <alley>, <village>, <cat>, <brick wall>, <woman>, <dress>, <laundry>, <line>, <buildings>, <bell tower>, <church>, <sky>
- Triplets: <alley, in, village>; <cat, in front of, woman>; <cat, sitting on, brick wall>; <woman, behind, cat>; <woman, wearing, dress>; <laundry, hanging from, line>; <line, between, buildings>; <bell tower, behind, woman>; <church, behind, buildings>; <bell tower, below, sky>; <bell tower, next to, church>

Exemplar 9 — Depth. *Prompt:* <On a frozen and desolate battlefield, a knight in gleaming silver armor stands in the foreground, raising a broadsword toward the ashen sky. Behind the knight, a row of soldiers holding wooden shields forms a defensive line across the icy plain. In the distant background, a dark fortress looms atop a jagged ridge, its turrets wreathed in black smoke. Scattered banners lie trampled on the snow-covered ground between the two armies.>

- Entities: <knight>, <broadsword>, <sky>, <armor>, <soldiers>, <shields>, <plain>, <fortress>, <ridge>, <banners>, <ground>
- Triplets: <broadsword, below, sky>; <knight, in front of, soldiers>; <knight, wearing, armor>; <knight, holding, broadsword>; <soldiers, behind, knight>; <soldiers, holding, shields>; <soldiers, standing on, plain>; <fortress, behind, soldiers>; <fortress, on top of, ridge>; <banners,

on, ground>; <banners, between, knight>;<banners, between, fortress>

Exemplar 10 — Depth. *Prompt:* <A bustling harbor scene unfolds under a cloudy afternoon sky. In the foreground, coils of thick rope and wooden barrels sit on the stone dock beside a moored fishing boat. Behind them, a crew of fishermen hauls a net full of silvery fish from the dark green water. In the far background, a massive cargo ship moves slowly across the horizon, its red hull contrasting against the grey ocean and distant fog.>

- Entities: <sky>, <rope>, <barrels>, <dock>, <fishing boat>, <fishermen>, <net>, <fish>, <water>, <cargo ship>, <horizon>
- Triplets: <rope, on, dock>; <barrels, on, dock>; <barrels, beside, fishing boat>; <rope, in front of, fishermen>; <fishermen, behind, barrels>; <fishermen, holding, net>; <fish, in, net>; <fishing boat, on, water>; <cargo ship, behind, fishermen>; <cargo ship, on, water>

Exemplar 11 — Centre. *Prompt:* <A grand ballroom glitters under the light of a crystal chandelier hanging from the ornate ceiling. In the center of the polished parquet floor, a couple in formal attire waltzes gracefully, their reflections shimmering below. To the left, a string quartet plays on a small raised stage draped in burgundy velvet. To the right, elegantly dressed guests stand beside tall marble columns, sipping champagne and watching the dancers with quiet admiration.>

- Entities: <ballroom>, <chandelier>, <ceiling>, <couple>, <reflections>, <floor>, <string quartet>, <stage>, <guests>, <columns>, <champagne>
- Triplets: <chandelier, hanging from, ceiling>; <chandelier, in, ballroom>; <chandelier, above, couple>; <couple, on, floor>; <couple, in, ballroom>; <reflections, below, couple>; <reflections, on, floor>; <couple, between, string quartet>; <couple, between, guests>; <string quartet, left of, couple>; <guests, right of, couple>; <string quartet, on, stage>; <guests, beside, columns>; <guests, holding, champagne>

Exemplar 12 — Centre. *Prompt:* <A quiet village square basks in warm golden afternoon sunlight. In the center, a moss-covered stone fountain gurgles softly, surrounded by pigeons pecking at crumbs on the weathered paving stones. To the left, a bakery with a striped awning displays loaves of bread and pastries behind its glass window. To the right, a flower shop overflows with buckets of sunflowers and lavender arranged on wooden crates outside its brightly painted door.>

- Entities: <square>, <fountain>, <pigeons>, <crumbs>, <paving stones>, <bakery>, <awning>, <bread>, <pastries>, <window>, <flower shop>, <sunflowers>, <lavender>, <crates>, <door>
- Triplets: <fountain, between, bakery>; <fountain, between, flower shop>; <bakery, left of, fountain>; <flower shop, right of, fountain>; <pigeons, around, fountain>; <pigeons, on, paving stones>; <crumbs, on, paving stones>; <crumbs, near, pigeons>; <bread, behind, window>; <pastries, behind, window>; <sunflowers, on, crates>; <lavender, on, crates>; <crates, beside, flower shop>; <door, of, flower shop>; <door, beside, crates>

Exemplar 13 — Centre. *Prompt:* <A spacecraft hangar stretches out in gleaming chrome and blue light. In the center, a sleek silver starfighter rests on a hydraulic landing platform, its cockpit canopy raised. To the left, a mechanic in an orange jumpsuit crouches beside a rack of plasma tools, inspecting a detached engine panel. To the right, stacks of metallic cargo containers are lined up near the massive hangar door, which opens onto the star-speckled void of deep space.>

- Entities: <hangar>, <starfighter>, <landing platform>, <cockpit>, <mechanic>, <jumpsuit>, <plasma tools>, <engine panel>, <cargo containers>, <hangar door>, <space>
- Triplets: <starfighter, in, hangar>; <starfighter, on, landing platform>; <starfighter, between, mechanic>;

<starfighter, between, cargo containers>; <mechanic, left of, starfighter>; <mechanic, wearing, jumpsuit>; <cargo containers, right of, starfighter>; <mechanic, beside, plasma tools>; <mechanic, next to, engine panel>; <cargo containers, near, hangar door>; <hangar door, in front of, space>

Exemplar 14 — Proximity-surface. *Prompt:* <On a weathered wooden pier at sunrise, a pelican sits on a barnacle-encrusted post, watching the gentle waves lap against the stilts below. Beside the post, a coil of frayed rope rests on the damp planks near an overturned red bucket. A small rowboat is tied to the pier, bobbing gently on the calm pink-tinged water. Seagulls stand on the railing nearby, their white feathers catching the first warm rays of morning light.>

- Entities: <pier>, <pelican>, <post>, <waves>, <stilts>, <rope>, <planks>, <bucket>, <rowboat>, <water>, <seagulls>, <railing>
- Triplets: <pelican, sitting on, post>; <post, on, pier>; <waves, below, pier>; <waves, hitting, stilts>; <stilts, below, pier>; <rope, beside, post>; <rope, on, planks>; <bucket, near, rope>; <bucket, on, planks>; <rowboat, tied to, pier>; <rowboat, in, water>; <seagulls, standing on, railing>; <railing, near, post>; <railing, on, pier>; <seagulls, near, pelican>

Exemplar 15 — Proximity-surface. *Prompt:* <A cozy mountain cabin interior glows with amber firelight on a snowy winter evening. A large grey cat lies on a thick wool rug beside the stone fireplace, its tail flicking lazily. Near the fireplace, a rocking chair draped with a patchwork quilt faces the crackling flames. On the mantelpiece above, a row of framed photographs sits next to a brass clock and two ceramic candlesticks. A steaming mug of cocoa rests on the side table.>

- Entities: <cabin>, <cat>, <rug>, <fireplace>, <fire>, <rocking chair>, <quilt>, <mantelpiece>, <photographs>, <clock>, <candlesticks>, <mug>, <side table>
- Triplets: <cat, lying on, rug>; <rug, beside, fireplace>; <rocking chair, near, fireplace>; <fire, in, fireplace>; <quilt, draped over, rocking chair>; <rocking chair, facing, fireplace>; <mantelpiece, above, fireplace>; <photographs, on, mantelpiece>; <clock, beside, photographs>; <candlesticks, beside, clock>; <mug, on, side table>; <side table, near, rocking chair>

Exemplar 16 — Proximity-surface. *Prompt:* <A cluttered artist's studio is bathed in soft north-facing window light. A half-finished oil painting rests on a wooden easel near the window, depicting a stormy seascape. Beside the easel, a stool holds a ceramic palette smeared with vivid blues and greens. Paint tubes and brushes are scattered on a low table next to the stool. On the floor near the door, a sleeping golden retriever curls up on a paint-spattered canvas dropcloth.>

- Entities: <studio>, <painting>, <seascape>, <easel>, <window>, <stool>, <palette>, <paint tubes>, <brushes>, <table>, <golden retriever>, <dropcloth>, <floor>, <door>
- Triplets: <painting, on, easel>; <easel, near, window>; <seascape, in, painting>; <stool, beside, easel>; <palette, on, stool>; <paint tubes, on, table>; <brushes, on, table>; <table, beside, stool>; <golden retriever, lying on, dropcloth>; <dropcloth, on, floor>; <golden retriever, near, door>

Exemplar 17 — Complex. *Prompt:* <A magical forest clearing teems with fantastical life at twilight. On the left, a unicorn drinks from a glowing blue stream that flows beneath a moss-draped stone arch. To the right, a group of pixies hovers above a cluster of giant luminescent mushrooms. In the center, an ancient wizard in purple robes sits on a fallen log, reading a thick leather-bound spellbook. Fireflies drift between the twisted trees surrounding the clearing.>

- Entities: <clearing>, <unicorn>, <stream>, <stone arch>, <pixies>, <mushrooms>, <wizard>, <robes>, <log>, <spellbook>, <fireflies>, <trees>

- Triplets: <unicorn, left of, wizard>; <unicorn, near, stream>; <stream, under, stone arch>; <pixies, right of, wizard>; <pixies, above, mushrooms>; <mushrooms, right of, wizard>; <wizard, between, unicorn>; <wizard, between, pixies>; <wizard, sitting on, log>; <wizard, wearing, robes>; <wizard, holding, spellbook>; <fireflies, between, trees>; <trees, surrounding, clearing>

Exemplar 18 — Complex. *Prompt:* <A chaotic pirate ship deck rolls on stormy seas under a dark, lightning-split sky. On the left, a parrot perches on the shoulder of a bearded captain who grips the wooden helm with both hands. To the right, a sailor climbs the rigging above stacked cannonballs on the deck. Behind the helm, a tattered black flag flaps wildly from the top of the main mast. Barrels of rum slide across the wet planks near the railing.>

- Entities: <pirate ship>, <seas>, <deck>, <sky>, <parrot>, <captain>, <helm>, <sailor>, <rigging>, <cannonballs>, <flag>, <mast>, <barrels>, <planks>, <railing>
- Triplets: <pirate ship, on, seas>; <deck, on, pirate ship>; <sky, above, pirate ship>; <parrot, on, captain>; <captain, left of, sailor>; <captain, holding, helm>; <sailor, right of, captain>; <sailor, climbing, rigging>; <sailor, above, cannonballs>; <cannonballs, on, deck>; <flag, behind, helm>; <flag, on top of, mast>; <mast, above, deck>; <barrels, on, planks>; <barrels, near, railing>; <helm, in front of, mast>; <railing, beside, deck>

Exemplar 19 — Complex. *Prompt:* <A sprawling underwater coral reef teems with vibrant marine life in crystal-clear tropical water. On the left, a sea turtle glides above a fan of bright orange coral near a rocky ledge. To the right, a school of silver fish swims beside a tall column of purple sponge. In the center, a scuba diver hovers between the turtle and the fish, shining a flashlight on a giant clam resting on the sandy ocean floor below.>

- Entities: <coral reef>, <water>, <sea turtle>, <orange coral>, <rocky ledge>, <fish>, <purple sponge>, <scuba diver>, <flashlight>, <giant clam>, <ocean floor>
- Triplets: <coral reef, in, water>; <sea turtle, left of, scuba diver>; <sea turtle, above, orange coral>; <orange coral, near, rocky ledge>; <fish, right of, scuba diver>; <fish, beside, purple sponge>; <scuba diver, between, sea turtle>; <scuba diver, between, fish>; <scuba diver, holding, flashlight>; <scuba diver, above, giant clam>; <giant clam, on, ocean floor>; <flashlight, facing, giant clam>

Exemplar 20 — Complex. *Prompt:* <An enchanted library stretches endlessly upward with spiraling staircases and floating bookshelves. On the left, a young witch in a pointed hat sits on a velvet armchair, leafing through a glowing grimoire. To the right, a spectral owl perches on the edge of a mahogany desk stacked with scrolls and inkwells. Above them, enchanted books flutter like birds between the towering shelves. A crystal orb rests on a pedestal in the center of the room.>

- Entities: <library>, <staircases>, <bookshelves>, <witch>, <hat>, <armchair>, <grimoire>, <owl>, <desk>, <scrolls>, <inkwells>, <books>, <shelves>, <crystal orb>, <pedestal>
- Triplets: <library, with, staircases>; <library, with, bookshelves>; <witch, left of, owl>; <witch, sitting on, armchair>; <witch, wearing, hat>; <witch, holding, grimoire>; <owl, right of, witch>; <owl, on the edge of, desk>; <scrolls, on, desk>; <inkwells, on, desk>; <books, above, witch>; <books, above, owl>; <books, between, shelves>; <shelves, in, library>; <crystal orb, on, pedestal>; <crystal orb, right of, witch>; <crystal orb, left of, owl>

B.3.2. Layout exemplars (entity bounding boxes)

Boxes are (x, y, w, h) top-left, normalized to $[0, 1]$, listed parallel to the entities of the matching exemplar id above.

Exemplar 1 — Horizontal.

entity	x	y	w	h
stone lantern	0.03	0.3	0.13	0.3
moss	0.0	0.56	0.22	0.08
maple tree	0.56	0.05	0.4	0.5
koi pond	0.5	0.55	0.48	0.25
leaves	0.62	0.57	0.11	0.08
gravel path	0.33	0.3	0.12	0.66
hedges	0.22	0.28	0.11	0.67

Exemplar 2 — Horizontal.

entity	x	y	w	h
dwarf	0.03	0.28	0.18	0.32
oak table	0.01	0.52	0.65	0.18
mug	0.1	0.47	0.07	0.08
elf	0.7	0.18	0.15	0.45
stone wall	0.68	0.0	0.32	0.72
back wall	0.0	0.0	1.0	0.45
parchment map	0.63	0.32	0.1	0.14
candle	0.6	0.22	0.06	0.1
fireplace	0.36	0.12	0.2	0.33
floor	0.0	0.45	1.0	0.55
smoky air	0.25	0.02	0.4	0.22

Exemplar 3 — Horizontal.

entity	x	y	w	h
highway	0.22	0.18	0.2	0.82
mountains	0.0	0.06	1.0	0.36
sky	0.0	0.0	1.0	0.14
gas station sign	0.01	0.22	0.14	0.38
cactus	0.57	0.5	0.07	0.28
adobe wall	0.4	0.26	0.14	0.16
pickup truck	0.44	0.38	0.22	0.16
crates	0.48	0.34	0.09	0.07
barbed wire	0.58	0.36	0.06	0.05
gravel shoulder	0.42	0.2	0.14	0.8

Exemplar 4 — Horizontal.

entity	x	y	w	h
plaza	0.0	0.35	1.0	0.65
market	0.0	0.1	1.0	0.9
woman	0.05	0.3	0.12	0.35
strawberries	0.18	0.42	0.08	0.06
straw hat	0.06	0.26	0.08	0.07
wooden stand	0.15	0.45	0.14	0.18
checkered cloth	0.16	0.48	0.12	0.06
old man	0.72	0.28	0.12	0.38
honey	0.6	0.44	0.06	0.06
folding table	0.56	0.46	0.14	0.14
children	0.38	0.42	0.14	0.22
busker	0.42	0.14	0.1	0.3
guitar	0.4	0.22	0.06	0.16
fountain	0.52	0.1	0.16	0.24
stalls	0.28	0.3	0.4	0.35

Exemplar 5 — Vertical.

entity	x	y	w	h
stained glass windows	0.1	0.1	0.8	0.35
patterns of vivid colored light	0.15	0.5	0.7	0.1
wooden pews	0.1	0.55	0.8	0.14
marble floor	0.0	0.82	1.0	0.18
chandelier	0.35	0.0	0.18	0.16
ceiling	0.0	0.0	1.0	0.2
candles	0.38	0.02	0.12	0.06
priest	0.4	0.62	0.12	0.28
robes	0.41	0.64	0.1	0.24
altar	0.3	0.7	0.3	0.12
lilies	0.34	0.67	0.06	0.06
golden cross	0.52	0.68	0.05	0.07
pigeons	0.68	0.36	0.12	0.08
rose window	0.62	0.12	0.22	0.24

Exemplar 6 — Vertical.

entity	x	y	w	h
oak tree	0.1	0.08	0.6	0.75
hillside	0.0	0.5	1.0	0.5
wildflowers	0.0	0.52	0.55	0.2
hawk	0.3	0.04	0.1	0.08
sky	0.0	0.0	1.0	0.12
treehouse	0.33	0.24	0.12	0.1
branches	0.14	0.1	0.52	0.38
rope ladder	0.37	0.34	0.04	0.34
ground	0.0	0.78	1.0	0.22
fox	0.3	0.68	0.12	0.1
leaves	0.26	0.74	0.2	0.06
boulder	0.44	0.66	0.1	0.1

Exemplar 7 — Vertical.

entity	x	y	w	h
waterfall	0.3	0.12	0.25	0.65
cliff	0.0	0.02	0.35	0.9
pool	0.2	0.75	0.65	0.25
ferns	0.05	0.35	0.15	0.15
rocks	0.03	0.45	0.18	0.12
bridge	0.15	0.04	0.7	0.1
river	0.2	0.02	0.6	0.5
hiker	0.42	0.0	0.08	0.14
mist	0.28	0.62	0.3	0.16

Exemplar 8 — Depth.

entity	x	y	w	h
alley	0.2	0.15	0.45	0.85
village	0.0	0.0	1.0	1.0
cat	0.25	0.6	0.2	0.22
brick wall	0.15	0.72	0.35	0.15
woman	0.3	0.32	0.14	0.3
dress	0.31	0.34	0.12	0.26
laundry	0.22	0.28	0.3	0.06
line	0.18	0.25	0.38	0.04
buildings	0.0	0.08	0.22	0.7
bell tower	0.42	0.05	0.12	0.22
church	0.3	0.1	0.2	0.18
sky	0.0	0.0	1.0	0.15

Exemplar 9 — Depth.

entity	x	y	w	h
knight	0.3	0.4	0.22	0.48
broadsword	0.42	0.3	0.06	0.28
sky	0.0	0.0	1.0	0.18
armor	0.31	0.42	0.2	0.44
soldiers	0.1	0.28	0.7	0.22
shields	0.12	0.3	0.66	0.1
plain	0.0	0.32	1.0	0.4
fortress	0.3	0.08	0.28	0.16
ridge	0.25	0.16	0.38	0.1
banners	0.18	0.22	0.5	0.08
ground	0.0	0.6	1.0	0.4

Exemplar 10 — Depth.

entity	x	y	w	h
sky	0.0	0.0	1.0	0.16
rope	0.16	0.66	0.1	0.08
barrels	0.22	0.62	0.12	0.14
dock	0.0	0.6	0.4	0.4
fishing boat	0.3	0.34	0.4	0.28
fishermen	0.36	0.36	0.22	0.22
net	0.62	0.38	0.14	0.18
fish	0.64	0.41	0.08	0.1
water	0.0	0.14	1.0	0.5
cargo ship	0.45	0.16	0.4	0.1
horizon	0.0	0.14	1.0	0.03

Exemplar 11 — Centre.

entity	x	y	w	h
ballroom	0.0	0.0	1.0	1.0
chandelier	0.38	0.02	0.18	0.18
ceiling	0.0	0.0	1.0	0.12
couple	0.38	0.38	0.18	0.3
reflections	0.36	0.68	0.22	0.08
floor	0.0	0.55	1.0	0.45
string quartet	0.04	0.32	0.2	0.22
stage	0.02	0.42	0.24	0.16
guests	0.72	0.3	0.18	0.32
columns	0.82	0.1	0.1	0.55
champagne	0.74	0.42	0.05	0.1

Exemplar 12 — Centre.

entity	x	y	w	h
square	0.0	0.0	1.0	1.0
fountain	0.36	0.3	0.2	0.28
pigeons	0.3	0.56	0.26	0.1
crumbs	0.34	0.62	0.1	0.05
paving stones	0.1	0.5	0.7	0.2
bakery	0.0	0.02	0.28	0.68
awning	0.2	0.28	0.14	0.06
bread	0.18	0.38	0.06	0.06
pastries	0.18	0.44	0.06	0.06
window	0.16	0.34	0.14	0.2
flower shop	0.72	0.02	0.28	0.68
sunflowers	0.66	0.42	0.07	0.1
lavender	0.6	0.44	0.07	0.08
crates	0.58	0.48	0.16	0.12
door	0.74	0.34	0.1	0.2

Exemplar 13 — Centre.

entity	x	y	w	h
hangar	0.0	0.0	1.0	1.0
starfighter	0.22	0.3	0.45	0.18
landing platform	0.18	0.46	0.52	0.08
cockpit	0.34	0.26	0.08	0.06
mechanic	0.08	0.38	0.07	0.14
jumpsuit	0.085	0.39	0.06	0.12
plasma tools	0.04	0.44	0.06	0.1
engine panel	0.15	0.4	0.08	0.08
cargo containers	0.72	0.24	0.12	0.3
hangar door	0.84	0.05	0.16	0.8
space	0.9	0.0	0.1	1.0

Exemplar 14 — Proximity-surface.

entity	x	y	w	h
pier	0.05	0.2	0.7	0.22
pelican	0.27	0.06	0.1	0.1
post	0.29	0.14	0.07	0.2
waves	0.05	0.46	0.9	0.5
stilts	0.15	0.36	0.4	0.12
rope	0.38	0.28	0.08	0.06
planks	0.08	0.3	0.6	0.1
bucket	0.48	0.27	0.06	0.07
rowboat	0.62	0.38	0.22	0.1
water	0.0	0.42	1.0	0.58
seagulls	0.14	0.14	0.1	0.06
railing	0.1	0.18	0.3	0.04

Exemplar 15 — Proximity-surface.

entity	x	y	w	h
cabin	0.0	0.0	1.0	1.0
cat	0.14	0.64	0.18	0.1
rug	0.08	0.62	0.32	0.16
fireplace	0.02	0.22	0.36	0.44
fire	0.08	0.34	0.22	0.24
rocking chair	0.44	0.4	0.22	0.38
quilt	0.45	0.42	0.2	0.34
mantelpiece	0.02	0.14	0.36	0.1
photographs	0.05	0.1	0.14	0.08
clock	0.21	0.1	0.07	0.08
candlesticks	0.3	0.1	0.06	0.08
mug	0.74	0.5	0.08	0.08
side table	0.7	0.46	0.16	0.18

Exemplar 16 — Proximity-surface.

entity	x	y	w	h
studio	0.0	0.0	1.0	1.0
painting	0.12	0.14	0.22	0.3
seascape	0.14	0.16	0.18	0.24
easel	0.1	0.1	0.26	0.5
window	0.0	0.02	0.16	0.5
stool	0.38	0.38	0.12	0.18
palette	0.4	0.34	0.08	0.06
paint tubes	0.54	0.46	0.08	0.05
brushes	0.52	0.42	0.06	0.04
table	0.5	0.42	0.16	0.14
golden retriever	0.68	0.66	0.2	0.12
dropcloth	0.62	0.64	0.3	0.16
floor	0.0	0.6	1.0	0.4
door	0.8	0.1	0.16	0.5

Exemplar 17 — Complex.

entity	x	y	w	h
clearing	0.1	0.1	0.8	0.8
unicorn	0.14	0.28	0.26	0.34
stream	0.0	0.0	0.12	1.0
stone arch	0.0	0.24	0.16	0.26
pixies	0.72	0.3	0.1	0.08
mushrooms	0.7	0.44	0.14	0.14
wizard	0.38	0.34	0.14	0.32
robes	0.39	0.36	0.12	0.28
log	0.34	0.62	0.22	0.08
spellbook	0.44	0.4	0.08	0.1
fireflies	0.2	0.14	0.5	0.12
trees	0.0	0.0	1.0	1.0

Exemplar 18 — Complex.

entity	x	y	w	h
pirate ship	0.05	0.22	0.9	0.55
seas	0.0	0.65	1.0	0.35
deck	0.08	0.5	0.84	0.22
sky	0.0	0.0	1.0	0.24
parrot	0.26	0.34	0.06	0.06
captain	0.2	0.32	0.14	0.3
helm	0.06	0.38	0.14	0.14
sailor	0.54	0.22	0.12	0.26
rigging	0.28	0.1	0.18	0.38
cannonballs	0.56	0.54	0.12	0.1
flag	0.18	0.06	0.08	0.1
mast	0.2	0.1	0.06	0.44
barrels	0.72	0.52	0.08	0.1
planks	0.1	0.52	0.8	0.1
railing	0.08	0.66	0.8	0.05

Exemplar 19 — Complex.

entity	x	y	w	h
coral reef	0.02	0.3	0.9	0.6
water	0.0	0.0	1.0	1.0
sea turtle	0.02	0.2	0.24	0.18
orange coral	0.04	0.44	0.18	0.16
rocky ledge	0.0	0.52	0.16	0.2
fish	0.66	0.2	0.26	0.16
purple sponge	0.76	0.36	0.1	0.3
scuba diver	0.36	0.22	0.16	0.34
flashlight	0.44	0.36	0.06	0.12
giant clam	0.32	0.66	0.22	0.12
ocean floor	0.0	0.76	1.0	0.24

Exemplar 20 — Complex.

entity	x	y	w	h
library	0.0	0.0	1.0	1.0
staircases	0.02	0.05	0.9	0.8
bookshelves	0.05	0.02	0.88	0.75
witch	0.08	0.36	0.14	0.3
hat	0.08	0.3	0.12	0.08
armchair	0.02	0.48	0.22	0.2
grimoire	0.16	0.42	0.08	0.1
owl	0.7	0.34	0.1	0.12
desk	0.62	0.46	0.26	0.16
scrolls	0.64	0.42	0.08	0.06
inkwells	0.76	0.43	0.05	0.05
books	0.2	0.08	0.5	0.16
shelves	0.14	0.04	0.66	0.28
crystal orb	0.4	0.48	0.1	0.1
pedestal	0.38	0.56	0.14	0.16

C. SGE-mini dataset splits

SpatialGenEval-mini draws on the SpatialGenEval corpus (approximately 1230 prompts). Each record's six-digit id is its SpatialGenEval running index, so the splits below are index sets into that corpus. **SGE-mini-1000** (1000 prompts) is the main evaluation set; **SGE-mini-val-100** (100 prompts) is a held-out validation set used for the cost-weight and guidance ablations (Appendix F). The two sets are disjoint.

SGE-mini-val-100 ids:

```
000005, 000014, 000024, 000036, 000038, 000040,
000058, 000062, 000063, 000082, 000102, 000109,
000114, 000124, 000125, 000152, 000167, 000181,
000183, 000192, 000205, 000210, 000215, 000216,
000231, 000237, 000243, 000256, 000257, 000267,
000277, 000279, 000290, 000292, 000315, 000317,
000321, 000335, 000339, 000360, 000361, 000391,
000403, 000404, 000428, 000436, 000452, 000461,
000473, 000501, 000502, 000503, 000553, 000557,
000561, 000562, 000572, 000577, 000578, 000580,
000590, 000593, 000600, 000608, 000611, 000612,
001055, 001131, 001132, 001133, 001134, 001135,
001136, 001137, 001138, 001139, 001140, 001144,
001145, 001151, 001152, 001153, 001154, 001155,
001156, 001157, 001158, 001159, 001169, 001170,
001177, 001180, 001182, 001184, 001185, 001190,
001201, 001203, 001204, 001205
```

SGE-mini-1000 ids (compact index ranges):

```
000001, 000003-000004, 000006-000011, 000013,
000015-000017, 000019-000023, 000025-000026,
000028-000035, 000037, 000039, 000041-000047,
000049-000053, 000055-000056, 000060-000061,
000064-000067, 000069-000075, 000078-000081,
000083-000085, 000087-000101, 000103-000108,
000110-000113, 000115-000122, 000126-000132,
000135-000151, 000153-000166, 000170-000174, 000176,
000179-000180, 000182, 000184-000191, 000193-000195,
000197-000204, 000206-000209, 000211-000214,
000217-000230, 000232-000236, 000238-000242, 000244,
000246-000255, 000258-000266, 000268-000276, 000278,
000280-000289, 000291, 000293-000301, 000303-000311,
000313, 000316, 000318-000320, 000322-000323,
000325-000330, 000332, 000334, 000336-000337,
000340-000359, 000362-000374, 000376-000382,
000384-000390, 000392, 000394-000402, 000405-000421,
000424-000427, 000430-000435, 000437, 000439-000446,
000448-000451, 000453-000460, 000462-000472,
000474-000477, 000479-000486, 000489-000496,
000498-000500, 000504, 000506-000523, 000525-000544,
000546-000552, 000556, 000558-000560, 000563-000571,
000573-000576, 000579, 000581, 000583, 000585-000589,
000591-000592, 000594-000596, 000598-000599,
000601-000607, 000609-000610, 000613-001010,
001012-001049, 001051-001054, 001056-001105,
001107-001119
```

D. FLUX single-stream versus dual-stream attention handling

FLUX.1-dev interleaves two transformer block types, and FOCAL's attention readout (Eq. 6) extracts the image-to-text attention submatrix from each in a slightly different way.

Block types. The nineteen dual-stream blocks keep the text and image streams separate, each with its own text query, key, and value projections. The thirty-eight single-stream blocks instead concatenate the text and image tokens into a single sequence and apply shared query, key, and value projections, realizing cross-attention by concatenation. FOCAL installs its capture hook on every block under one contiguous index space, with the dual-stream blocks indexed first and the single-stream blocks following.

Locating the text tokens. Both block types extract the same image-to-text slice; they differ only in how the text length is recovered at the attention call. In the dual-stream blocks the text stream is available separately, so its length is known directly. In the single-stream blocks the text and image tokens are concatenated before attention, so the text length is supplied once per denoising step and reused when slicing.

Readout. Given the text length, both paths take the image-query rows against the text-key columns, average over attention heads, and transpose so that each column is the spatial attention map of one text token. These per-block maps are then averaged over a configurable band of blocks to form Eq. 6.

Block band. The set of guided blocks is a modeling choice: one may guide only the single-stream blocks (as DreamRenderer does), only the dual-stream blocks, or a band spanning the last dual-stream blocks and the first single-stream blocks, which is our default. The FOCUS attention-binding objective is framed around the dual-stream text projections, whereas the single-stream blocks expose only the shared, concatenated attention, so the two stream types carry text grounding in structurally different ways.

E. Scene-graph parser ablation

We compare five parsers by **spatial-relation recall** on the DiscoSG-DS test set (100 examples) over **18 spatial relation groups** adapted from the VSR 66-relation taxonomy. The two LLM parsers use 20-example few-shot ICL (the prompt of Appendix B.1); manner verbs are folded into spatial groups ("sitting on" → on, "hang above" → above).

Why recall. For layout guidance a *missed* relation leaves an entity pair unguided and default-placed, whereas an *extra* non-spatial relation is a possibly-wrong constraint that correct ones override. The study therefore optimises recall. It documents the trade-off: LLM parsers emit ∼12 spatial triplets/example vs ∼4 for DiscoSG, giving higher recall but lower precision (∼0.14 vs ∼0.60).

Matching rules (5). group-compatibility (predicate maps to the same or a subsuming group); bidirectional substring entity match; bidirectional argument order (active/-passive); meta-entity exclusion (background/foreground never an object); greedy 1:1 (no double-count).

The five parsers. GPT-5.1 (LLM, 20-shot), DeepSeek-V3.2 (LLM, 20-shot), DiscoSG Refiner (fine-tuned), FACTUAL-T5 (fine-tuned), spaCy (rule-based).

Surface predicates for the 18 relation groups.

Relation group	Surface predicates
on	on, on top of, on the surface of, upon, atop, on face of, on far side of, sit upon, perch atop
covered by / wearing	covered by, cover, cover with, wearing
in / inside	in, inside, within, into, in end of, in corner of, in center of, on middle of
near / next to	near, next to, close to, beside, by, alongside, adjacent to, among, on side of, on edge of, on both sides of, on either side of, at, sit at, stand at, stop at, sit by, stand by
above / over	above, over, at the top of, overlook, over side of
around / surrounding	around, surrounding, enclosed by, surround, wrap around, tie around, enclose
holding / carrying	holding, carrying, hold, carry
below / under	below, under, beneath, underneath, at the bottom of, on bottom of, at base of, at foot of
against / along	against, along, across, glide across, move across
left of	left of, to the left of, on the left of, at the left side of, at the left of, on the left side of
behind	behind, in back of, on the back of, on back of
in front of	in front of, ahead of, on front of
facing / orientation	facing, facing away from, parallel to, perpendicular to, across from, face, opposite of, lay parallel to, towards, toward, walk towards, run towards, stand towards, dive towards
right of	right of, to the right of, on the right of, at the right side of, on the right side of
center / middle	in center of, in the middle of, on middle of, center of
far from	far from, far away from, away from, beyond, move away from
through	through
attached / connected	attached to, connected to, attach to, connect to, mount on, attach, connect, attach with

Posture and manner predicates (for example, lying, sitting, and standing on) are folded into the nearest group.

E.1. Headline micro / macro recall

Parser	Type	Micro	Macro
GPT-5.1 (adopted)	LLM (20-shot)	**0.682**	**0.550**
DeepSeek-V3.2	LLM (20-shot)	0.628	0.485
DiscoSG Refiner	fine-tuned	0.583	0.546
FACTUAL-T5	fine-tuned	0.171	0.082
spaCy	rule-based	0.109	0.109

Few-shot GPT-5.1 leads on both micro and macro recall, with its largest margin on micro. DeepSeek-V3.2 is a close second and wins a few groups. DiscoSG Refiner ties GPT on macro (0.546 vs 0.550) and is strongest on rare relations (against/along, through, attached), but trails sharply on micro. FACTUAL-T5 and spaCy under-generate badly on this distribution.

E.2. Per-relation spatial recall (18 groups)

Bold = best in row. "Sup." is the support (number of gold relations in that group).

Relation group	Sup.	GPT-5.1	DeepSeek	DiscoSG	FACT-T5	spaCy
on	109	**0.761**	0.734	0.596	0.202	0.183
covered by / wearing	85	**0.800**	0.729	0.765	0.376	0.059
in / inside	39	**0.692**	0.615	0.359	0.077	0.051
near / next to	34	0.618	**0.647**	0.559	0.029	0.059
above / over	25	**0.680**	0.640	0.400	0.080	0.160
around / surrounding	24	**0.708**	0.375	0.583	0.083	0.042
holding / carrying	24	**0.667**	0.542	**0.667**	0.167	0.083
below / under	15	**0.533**	**0.533**	0.400	0.067	0.133
against / along	14	0.143	0.214	**0.714**	0.286	0.214
left of	10	**1.000**	0.800	0.700	0.100	0.000
behind	9	0.556	**0.778**	0.667	0.000	0.333
in front of	8	0.250	**0.500**	0.375	0.000	0.000
facing / orientation	7	**0.571**	0.286	0.286	0.000	0.143
right of	6	0.833	**1.000**	0.500	0.000	0.000
center / middle	4	0.000	0.000	**0.250**	0.000	0.000
far from	4	**0.250**	0.000	0.000	0.000	0.000
through	3	0.333	0.333	**1.000**	0.000	0.000
attached / connected	2	0.500	0.000	**1.000**	0.000	0.500

GPT-5.1 leads the high-support rows decisively (on +28%, in +93%, above +70%, left-of +43% vs DiscoSG). Low-support rows (Sup. ¡= 4: center, far from, through, attached) are noisy, so conclusions weight toward high-support relations. We adopt GPT-5.1 and, for consistency, use the same model for the layout planner.

F. FOCAL cost-weight ablation

We grid-search the three FOCAL cost weights of Eq. 7 on SGE-mini-val-100 (100 prompts; Appendix C), holding the parser, the layout, and the guidance strength (λ_{env}=500) fixed. Metrics are multiple-choice VQA accuracies on the five SGE-mini axes (Object, Position, Orientation, Proximity, Occlusion); **spatial** is the mean of Position, Orientation, Proximity, and Occlusion, and **avg** is the mean of all five. All 84 evaluated configurations are listed below, sorted by average accuracy. The adopted configuration (w_{d}=0.2, w_{t}=2.0, w_{c}=10.0) is marked with a star and attains the best average (0.724) and spatial (0.685) accuracy.

	w_{d}	w_{t}	w_{c}	Object	Position	Orientation	Proximity	Occlusion	spatial	avg
⋆	0.2	2.0	10.0	0.88	0.76	0.61	0.78	0.59	0.685	0.724
	0.2	0.5	9.0	0.88	0.77	0.65	0.71	0.56	0.672	0.714
	0.35	0.5	10.0	0.87	0.73	0.59	0.71	0.61	0.660	0.702
	0.2	1.0	9.0	0.82	0.74	0.59	0.78	0.57	0.670	0.700
	0.2	2.0	7.0	0.83	0.75	0.63	0.74	0.55	0.667	0.700
	0.35	2.0	7.0	0.85	0.79	0.63	0.69	0.54	0.662	0.700
	0.2	2.0	9.0	0.86	0.72	0.64	0.71	0.57	0.660	0.700
	0.2	4.0	10.0	0.84	0.78	0.58	0.71	0.55	0.655	0.692
	0.35	1.0	10.0	0.87	0.72	0.63	0.69	0.55	0.647	0.692
	0.05	2.0	10.0	0.84	0.77	0.61	0.71	0.52	0.652	0.690
	0.2	0.5	10.0	0.83	0.75	0.62	0.71	0.53	0.652	0.688
	0.2	0.5	15.0	0.86	0.73	0.63	0.67	0.54	0.642	0.686
	0.2	4.0	15.0	0.86	0.73	0.58	0.70	0.56	0.642	0.686
	0.05	1.0	10.0	0.87	0.75	0.60	0.70	0.51	0.640	0.686
	0.35	1.0	5.0	0.84	0.76	0.63	0.71	0.48	0.645	0.684
	0.2	0.5	8.0	0.82	0.69	0.63	0.76	0.51	0.647	0.682
	0.2	1.0	20.0	0.83	0.76	0.62	0.69	0.50	0.642	0.680
	0.05	0.5	10.0	0.81	0.74	0.62	0.71	0.51	0.645	0.678
	0.2	1.0	7.0	0.83	0.74	0.62	0.64	0.56	0.640	0.678
	0.2	4.0	9.0	0.84	0.75	0.58	0.69	0.53	0.637	0.678
	0.2	2.0	15.0	0.81	0.73	0.62	0.69	0.53	0.642	0.676
	0.2	1.0	8.0	0.82	0.71	0.60	0.73	0.52	0.640	0.676
	0.2	1.0	10.0	0.84	0.74	0.61	0.71	0.48	0.635	0.676
	0.05	1.0	7.0	0.86	0.75	0.63	0.68	0.46	0.630	0.676
	0.2	1.0	5.0	0.82	0.74	0.64	0.73	0.44	0.637	0.674
	0.2	2.0	8.0	0.89	0.73	0.61	0.66	0.48	0.620	0.674
	0.35	2.0	10.0	0.83	0.73	0.59	0.71	0.50	0.632	0.672
	0.2	0.5	20.0	0.74	0.78	0.64	0.64	0.55	0.652	0.670
	0.35	0.5	7.0	0.85	0.72	0.61	0.70	0.47	0.625	0.670
	0.2	4.0	8.0	0.79	0.76	0.58	0.69	0.52	0.637	0.668
	0.05	2.0	7.0	0.87	0.71	0.63	0.69	0.44	0.618	0.668
	0.05	0.5	5.0	0.79	0.76	0.65	0.69	0.44	0.635	0.666
	0.2	1.0	15.0	0.80	0.70	0.61	0.71	0.51	0.632	0.666
	0.05	2.0	5.0	0.83	0.73	0.62	0.67	0.48	0.625	0.666
	0.35	0.5	2.0	0.84	0.77	0.64	0.69	0.39	0.623	0.666
	0.35	1.0	7.0	0.81	0.71	0.63	0.65	0.52	0.627	0.664
	0.05	0.5	7.0	0.82	0.73	0.58	0.70	0.49	0.625	0.664
	0.35	2.0	5.0	0.82	0.78	0.60	0.70	0.42	0.625	0.664
	0.2	2.0	4.0	0.83	0.73	0.60	0.69	0.47	0.623	0.664
	0.2	0.5	7.0	0.84	0.70	0.63	0.71	0.44	0.620	0.664
	0.35	0.5	5.0	0.84	0.74	0.62	0.70	0.42	0.620	0.664
	0.2	4.0	20.0	0.74	0.71	0.60	0.75	0.51	0.642	0.662

w_d	w_t	w_c	Object	Position	Orientation	Proximity	Occlusion	spatial	avg
0.2	2.0	5.0	0.80	0.74	0.60	0.73	0.44	0.627	0.662
0.05	1.0	4.0	0.82	0.74	0.62	0.68	0.45	0.623	0.662
0.2	2.0	20.0	0.75	0.73	0.60	0.69	0.53	0.637	0.660
0.2	2.0	3.0	0.80	0.71	0.66	0.69	0.44	0.625	0.660
0.2	1.0	3.0	0.82	0.76	0.61	0.69	0.42	0.620	0.660
0.35	1.0	4.0	0.84	0.77	0.62	0.67	0.40	0.615	0.660
0.2	1.0	4.0	0.78	0.72	0.61	0.70	0.48	0.627	0.658
0.05	2.0	3.0	0.80	0.74	0.63	0.70	0.41	0.620	0.656
0.2	4.0	30.0	0.74	0.69	0.62	0.69	0.53	0.632	0.654
0.2	2.0	2.0	0.79	0.74	0.67	0.68	0.39	0.620	0.654
0.2	0.5	4.0	0.81	0.73	0.60	0.71	0.42	0.615	0.654
0.35	2.0	2.0	0.86	0.71	0.61	0.64	0.45	0.603	0.654
0.05	1.0	5.0	0.79	0.73	0.63	0.68	0.43	0.618	0.652
0.35	1.0	2.0	0.80	0.74	0.64	0.65	0.43	0.615	0.652
0.05	0.5	3.0	0.78	0.74	0.63	0.72	0.38	0.618	0.650
0.2	0.5	5.0	0.82	0.68	0.62	0.72	0.41	0.608	0.650
0.2	2.0	30.0	0.69	0.67	0.61	0.69	0.58	0.637	0.648
0.2	0.5	30.0	0.74	0.74	0.58	0.65	0.53	0.625	0.648
0.35	0.5	4.0	0.79	0.72	0.60	0.67	0.46	0.613	0.648
0.05	2.0	4.0	0.80	0.73	0.59	0.69	0.43	0.610	0.648
0.05	0.5	4.0	0.81	0.76	0.60	0.69	0.38	0.608	0.648
0.35	2.0	4.0	0.83	0.71	0.59	0.70	0.40	0.600	0.646
0.05	1.0	2.0	0.79	0.72	0.65	0.67	0.39	0.608	0.644
0.35	2.0	3.0	0.79	0.71	0.63	0.68	0.40	0.605	0.642
0.35	0.5	3.0	0.80	0.74	0.64	0.66	0.37	0.603	0.642
0.35	1.0	3.0	0.80	0.74	0.63	0.67	0.37	0.603	0.642
0.2	1.0	2.0	0.74	0.75	0.66	0.66	0.39	0.615	0.640
0.2	0.5	3.0	0.76	0.76	0.63	0.63	0.41	0.608	0.638
0.05	1.0	3.0	0.78	0.72	0.60	0.66	0.43	0.603	0.638
0.2	0.5	2.0	0.80	0.73	0.62	0.66	0.36	0.593	0.634
0.35	2.0	0.5	0.77	0.70	0.61	0.69	0.39	0.598	0.632
0.2	2.0	0.5	0.75	0.74	0.65	0.65	0.36	0.600	0.630
0.05	2.0	2.0	0.73	0.77	0.63	0.66	0.35	0.603	0.628
0.2	1.0	30.0	0.68	0.72	0.60	0.62	0.51	0.613	0.626
0.35	0.5	0.5	0.76	0.70	0.66	0.64	0.35	0.588	0.622
0.05	0.5	2.0	0.76	0.73	0.61	0.62	0.38	0.585	0.620
0.2	1.0	0.5	0.74	0.72	0.66	0.65	0.31	0.585	0.616
0.35	1.0	0.5	0.74	0.71	0.59	0.67	0.34	0.578	0.610
0.2	0.5	0.5	0.76	0.71	0.62	0.60	0.32	0.562	0.602
0.05	2.0	0.5	0.70	0.74	0.63	0.61	0.32	0.575	0.600
0.05	1.0	0.5	0.72	0.69	0.64	0.62	0.31	0.565	0.596
0.05	0.5	0.5	0.64	0.66	0.61	0.65	0.36	0.570	0.584

3

Preliminaries - Extended

This chapter develops in full the material that Chapter 2 presents in condensed form. Section 3.1 reviews rectified-flow sampling and the multi-modal diffusion transformer (MM-DiT) that parameterizes the velocity field, and Section 3.2 reviews the stochastic-optimal-control fundamental theory.

3.1. Rectified-Flow Sampling on MM-DiT

Modern transformer-based text-to-image (T2I) generators such as Stable Diffusion 3.5 and FLUX [11, 5] synthesize an image by integrating an ordinary differential equation (ODE) whose right-hand side is a velocity field trained under the *flow matching* (FM) framework [20, 21, 1], in the latent space of a pretrained variational autoencoder [15, 29]. The encoder $\mathcal{E} : \mathbb{R}^{3\times H_{\text{img}}\times W_{\text{img}}} \to \mathbb{R}^{C\times H\times W}$ maps an image to a latent z that retains its perceptual content, the decoder $\mathcal{D}$ reconstructs an image from it, and both are held fixed throughout this thesis; all generative dynamics described below act on the latent, which has tensor shape (C, H, W) and which we equivalently view as a single point in $\mathbb{R}^d$ with $d = CHW$ when writing equations.

Probability paths and the generative ODE. Let $\pi_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ be a standard Gaussian noise distribution over latents and let π_1 be the distribution of latents corresponding to realistic images, i.e. the empirical distribution of $\mathcal{E}(x)$ over training images x. A rectified-flow T2I model synthesizes an image by integrating the ODE

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} \;=\; v_\theta(X_t, t, c), \qquad X_0 \sim \pi_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \tag{3.1}$$

over $t \in [0, 1]$ in the frozen VAE latent space, where c is the text conditioning and v_θ is the trained velocity field. The solution path $\{X_t\}_{t\in[0,1]}$ has the correct marginals: at every intermediate time t, the distribution of X_t matches a family $\{\pi_t\}$ that interpolates between π_0 at $t = 0$ and π_1 at $t = 1$, following the convention of Lipman et al. [20] in which the base noise sits at $t = 0$ and the data at $t = 1$. Inference is therefore the forward-time Euler integration of Equation (3.1) from $t = 0$ to $t = 1$, after which the terminal latent X_1 is VAE-decoded to the output image $\mathcal{D}(X_1)$. At each state $X_t \in \mathbb{R}^d$ and time t, the velocity $v_\theta(X_t, t, c) \in \mathbb{R}^d$ is the tangent vector along which the trajectory evolves, and it carries the same (C, H, W) spatial structure as the latent itself.

Reference paths and rectified flow. To avoid simulating the ODE during training, FM specifies an explicit reference path between samples of the two endpoint distributions. Given a coupling $(\bar{X}_0, \bar{X}_1) \sim \pi_0 \times \pi_1$, the path is the linear interpolation [20, 1]

$$\bar{X}_t \;=\; \beta_t\, \bar{X}_0 \;+\; \alpha_t\, \bar{X}_1, \qquad t \in [0, 1], \tag{3.2}$$

with smooth scalar schedules α_t, β_t satisfying $\alpha_0 = 0$, $\beta_0 = 1$, $\alpha_1 = 1$, $\beta_1 = 0$, α_t strictly increasing and β_t strictly decreasing. The bar distinguishes these *reference-path* quantities, defined for one fixed coupled pair $(\bar{X}_0, \bar{X}_1)$, from the unbarred state X_t of the generative ODE in Equation (3.1): $\bar{X}_t$ is a known closed-form interpolation between two specific endpoints, used only to define the training target, whereas X_t is the actual sample produced at inference by integrating the learned field v_θ and is tied to no particular endpoints; the two share the same marginal distribution π_t at every t by construction. The pathwise derivative of the reference path,

$$u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1) \;:=\; \frac{\mathrm{d}\bar{X}_t}{\mathrm{d}t} \;=\; \dot{\beta}_t\, \bar{X}_0 \;+\; \dot{\alpha}_t\, \bar{X}_1, \tag{3.3}$$

is the conditional velocity along it. The instance used by SD3.5 and FLUX is *rectified flow* (RF) [21], obtained with the linear schedule $(\alpha_t, \beta_t) = (t, 1-t)$, under which the reference path $\bar{X}_t = (1-t)\,\bar{X}_0 + t\,\bar{X}_1$ is a straight line from the noise sample at $t = 0$ to the data sample at $t = 1$ and Equation (3.3) reduces to the constant-in-time velocity $u_t = \bar{X}_1 - \bar{X}_0$. Only the *conditional* reference paths are straight, however; the *marginal* velocity field that the network learns to approximate, defined next, is in general a non-trivial function of X_t and t.

Training Objective. The marginal velocity field is the conditional expectation of Equation (3.3) over all couplings whose reference path passes through X_t at time t,

$$u_t(X_t) \;:=\; \mathbb{E}_{(\bar{X}_0, \bar{X}_1)\sim\pi_0\times\pi_1}\Big[\, u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1) \;\Big|\; \bar{X}_t = X_t \,\Big], \tag{3.4}$$

which is intractable, because evaluating it would require sampling from the marginal density of $\bar{X}_t$, so flow matching instead trains v_θ against the conditional flow-matching (CFM) loss [20] (suppressing the conditioning c for clarity)

$$\mathcal{L}_{\mathrm{CFM}}(\theta) \;=\; \mathbb{E}_{t\sim\pi(t)}\, \mathbb{E}_{\bar{X}_0\sim\pi_0,\, \bar{X}_1\sim\pi_1}\Big[\, \big\| v_\theta(\bar{X}_t, t) - u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1) \big\|_2^2 \,\Big], \tag{3.5}$$

where t is drawn from a sampling density $\pi(t)$ on $[0, 1]$ and $(\bar{X}_0, \bar{X}_1)$ from the coupling. It can be shown [20, Theorem 2] that $\nabla_\theta \mathcal{L}_{\mathrm{CFM}} = \nabla_\theta \mathcal{L}_{\mathrm{FM}}$, where

$$\mathcal{L}_{\mathrm{FM}}(\theta) \;=\; \mathbb{E}_{t\sim\pi(t)}\, \mathbb{E}_{X_t\sim p_t}\big[\, \| v_\theta(X_t, t) - u_t(X_t) \|_2^2 \,\big] \tag{3.6}$$

is the intractable marginal loss, with p_t the marginal density of the ODE state X_t at time t, so the two objectives differ only in their per-sample residuals and the network learns the marginal field even though every gradient step sees a single conditional pair.

Sampling. At inference an initial state $X_0 \sim \pi_0$ is drawn and Equation (3.1) is integrated forward on a grid $0 = t_0 < t_1 < \cdots < t_N = 1$ with a first-order Euler scheme,

$$X_{t_{i+1}} \;=\; X_{t_i} \;+\; (t_{i+1} - t_i)\, v_\theta(X_{t_i}, t_i, c), \qquad i = 0, \ldots, N-1, \tag{3.7}$$

after which $X_{t_N} \approx X_1$ is VAE-decoded; the dominant cost is the N forward evaluations of v_θ.

Classifier-Free Guidance. To strengthen prompt adherence, T2I flow-matching models are sampled with classifier-free guidance (CFG) [14], which at each step replaces the conditional prediction by an extrapolation between the conditional and unconditional outputs,

$$\tilde{v}_\theta(X_t, t, c) \;=\; v_\theta(X_t, t, \emptyset) \;+\; s \cdot \big(v_\theta(X_t, t, c) - v_\theta(X_t, t, \emptyset)\big), \tag{3.8}$$

with $\emptyset$ the null conditioning obtained by encoding an empty prompt and $s \geq 1$ the guidance scale. CFG strengthens conditional generation on average but is agnostic to spatial structure: it pushes the velocity

toward higher $\log p(c \mid X_t)$ with no mechanism to bias the trajectory toward a *specified* layout, the gap that FOCAL fills.

The MM-DiT Velocity Network. The velocity field v_θ is parameterized by a multi-modal diffusion transformer (MM-DiT) [11, 5] built on the diffusion-transformer backbone of Peebles and Xie [22]. The network receives the current latent X_t, the time t, and a text conditioning c from one or more frozen text encoders [26, 27]; the latent is patchified into image tokens $\mathbf{X}^{(\text{img})} \in \mathbb{R}^{N_{\text{img}} \times D}$ and the encoders produce text tokens $\mathbf{X}^{(\text{txt})} \in \mathbb{R}^{N_{\text{txt}} \times D}$, which are processed by a stack of K joint-attention blocks. In each block the concatenated sequence $[\mathbf{X}^{(\text{img})}; \mathbf{X}^{(\text{txt})}]$ is passed through a single self-attention operator with modality-specific query, key, and value projections but attention computed jointly over the union of the two streams, so information flows bidirectionally between modalities at every block, in contrast to the one-way text-to-image cross-attention of earlier UNet latent-diffusion models [29, 24]. Each block produces a single softmax matrix indexed by all $(N_{\text{img}} + N_{\text{txt}})$ token positions, and its *image-to-text submatrix* gives, for each text token, a distribution over image positions, equivalently a spatial map over the image grid; these per-token spatial maps are the differentiable signal that the running cost of FOCAL reads and differentiates through.

3.2. Stochastic Optimal Control for FM-SDE

We now nudge the rectified-flow dynamics of Section 3.1 so that the generated image satisfies a spatial-layout constraint supplied externally as a set of bounding boxes, while keeping the controlled trajectory close to the base flow. Stochastic optimal control (SOC) provides a principled framework for this inference-time intervention [36, 12, 4, 31]: an additive control term enters the dynamics, and the trade-off between satisfying the constraint and staying close to the base sampler is encoded as a cost function that the control minimizes. The specialization to our spatial cost and the rectified-flow schedule is deferred to the methodology in Chapter 2.

Controlled Dynamics and Cost. Following [10, 4], a time-dependent control $u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ turns the ODE (3.1) into the controlled stochastic differential equation (SDE)

$$\mathrm{d}X_t^u \;=\; \big(b(X_t^u, t) + \sigma(t)\, u(X_t^u, t)\big)\,\mathrm{d}t \;+\; \sigma(t)\,\mathrm{d}B_t, \qquad X_0^u \sim \pi_0, \tag{3.9}$$

where B_t is a standard Brownian motion in $\mathbb{R}^d$, $\sigma : [0,1] \to \mathbb{R}_{\geq 0}$ is a scalar diffusion schedule, and the FM base drift b equals the trained velocity plus a score correction that cancels the diffusion term at the level of marginals, so that for $u \equiv 0$ the SDE samples the same path marginals as the deterministic FM ODE for every σ [4, Eq. 4]. We choose u to balance two requirements: the trajectory should remain close to the base sampler, so the model keeps producing realistic images rather than drifting off-manifold, and it should be steered toward states that satisfy our objective. The first requirement has an exact form: by Girsanov's theorem the control penalty $\frac{1}{2}\|u\|_2^2$, integrated over time, equals the Kullback–Leibler divergence of the controlled path measure from the base one [3], so penalizing it is mathematically a KL regularization against the base sampler. The second requirement is encoded by a running cost $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}$ scoring how well intermediate states align with the layout target and a terminal cost $g : \mathbb{R}^d \to \mathbb{R}$ on the final state, such as a CLIP [26] or preference score; we set $g \equiv 0$ throughout following [4] and encode the spatial objective entirely in f. Together these give the cost function

$$J(u) \;=\; \mathbb{E}_{X^u \sim \mathbb{P}^u}\left[\int_0^1 \Big(\tfrac{1}{2}\,\|u(X_t^u, t)\|_2^2 \;+\; f(X_t^u, t)\Big)\,\mathrm{d}t \;+\; g(X_1^u)\right], \tag{3.10}$$

where the expectation is over the controlled trajectory $X^u \sim \mathbb{P}^u$ induced by Equation (3.9), and whose minimizer is the control that reduces the trajectory cost while staying close to the base path measure.

Optimal Control via Pontryagin. Minimizing Equation (3.10) over the whole control trajectory subject to the dynamics (3.9) is an infinite-dimensional problem, but the Pontryagin minimum principle [25, 19] reduces it to a pointwise condition by introducing an auxiliary backward process. The Hamiltonian collects the running-cost integrand and the inner product of the costate $a(t) \in \mathbb{R}^d$ with the drift,

$$\mathcal{H}(x,u,a,t) \;=\; \tfrac{1}{2}\|u\|_2^2 \;+\; f(x,t) \;+\; a(t)^\top\big(b(x,t)+\sigma(t)\,u\big), \tag{3.11}$$

where the costate is the gradient of the value function $V(x,t)$, the smallest total cost still attainable from state x at time t, so $a(t)=\nabla_x V$ measures the sensitivity of the future cost to the current state. The Hamiltonian is strictly convex in u (its u-Hessian is $\mathbf{I}_d$, every other term being at most linear in u), so the first-order condition $\nabla_u\mathcal{H}=0$ gives the unique minimizer $u^\star(x,t)=-\sigma(t)\,a(t)$, proportional to the costate and pointing toward decreasing expected future cost. The costate satisfies the backward adjoint ODE

$$\frac{\mathrm{d}a(t)}{\mathrm{d}t} \;=\; -\nabla_x b(X_t^u,t)^\top a(t) \;-\; \nabla_x f(X_t^u,t), \qquad a(1) \;=\; \nabla_x g(X_1^u), \tag{3.12}$$

which together with the forward dynamics (3.9) and the optimality condition forms a coupled forward–backward boundary-value problem.

Single-pass approximation. Solving that system exactly during sampling would require a full backward solve after every forward step and is computationally prohibitive, so following the test-time controller of FOCUS [4, Eq. 11] we make two approximations. First, the term $\nabla_x b^\top a$ in Equation (3.12) is dropped, assuming the local sensitivity of the base velocity to state perturbations is small relative to the running-cost gradient, a standard online-control approximation [13]; with $g\equiv 0$ this leaves $a(t)=\int_t^1 \nabla_x f(X_s^u,s)\,\mathrm{d}s$. Second, the integrand is frozen at its current value over $[t,1]$, a left-Riemann approximation that yields $a(t)\approx(1-t)\,\nabla_x f(X_t^u,t)$ and hence the single-pass optimal control

$$u_t^\star \;\approx\; -\sigma(t)\,(1-t)\,\nabla_x f\big(X_t^u,t\big). \tag{3.13}$$

From control to velocity. The control acts on the SDE drift b, but at inference, we run the deterministic ODE, so we recast it as a velocity correction. An SDE and an ODE produce the same marginal distribution of X_t at every t when the deterministic velocity equals the drift minus the score term $\frac{1}{2}\sigma^2(t)\,\nabla_x \log p_t$, a consequence of the Fokker–Planck equation [28, 30]. For the uncontrolled SDE this rule must return the model's own velocity, so $b=v_\theta+\frac{1}{2}\sigma^2(t)\,\nabla_x\log p_t$, and applying it to the controlled drift $b+\sigma(t)\,u^\star$ with the score shift $\nabla_x\log p_t^u-\nabla_x\log p_t=u_t^\star/\sigma(t)$ [3, 10] and Equation (3.13) gives the controlled velocity

$$v_t^u \;=\; v_\theta(X_t,t,c) \;+\; \tfrac{1}{2}\sigma(t)\,u_t^\star \;=\; v_\theta(X_t,t,c) \;-\; \eta(t)\,\nabla_x f(X_t^u,t), \qquad \eta(t)=\tfrac{1}{2}\sigma^2(t)\,(1-t). \tag{3.14}$$

The prefactor $\eta(t)$ is the step-size weight that the SOC derivation imposes on the gradient-style velocity correction.

The Memoryless Schedule. Equation (3.14) leaves $\sigma(t)$ unspecified, but it is not a free choice: the optimal control reweights each image's likelihood by $e^{-\int_0^1 f\,\mathrm{d}t}$, so for the starting noise to remain Gaussian after reweighting, and hence for the controlled process to still be sampled from $X_0\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, the initial noise X_0 and the data X_1 must be independent under the base process. Domingo-Enrich et al. [10, Prop. 1, Thm. 1] prove that within the FM–SDE family a unique schedule enforces this *memoryless* property,

$$\sigma_{\mathrm{mem}}^2(t) \;=\; 2\,\beta_t\Big(\tfrac{\dot\alpha_t}{\alpha_t}\beta_t-\dot\beta_t\Big), \tag{3.15}$$

which for the rectified-flow interpolant $(\alpha_t,\beta_t)=(t,1-t)$ evaluates to $\sigma_{\mathrm{mem}}^2(t)=2(1-t)/t$, and substituting into the weight of Equation (3.14) gives the closed-form rectified-flow weight

$$\eta(t) \;=\; \tfrac{1}{2}\,\sigma_{\mathrm{mem}}^2(t)\,(1-t) \;=\; \frac{(1-t)^2}{t}. \tag{3.16}$$

The weight diverges as $t \to 0^+$, giving the strong intervention needed when the sample is mostly noise, and decays to zero at $t = 1$, leaving the late content-committing steps untouched. The schedule is required only for the SOC derivation and its guarantees; at inference, one is free to integrate the controlled velocity with any sampler, including the deterministic $\sigma \equiv 0$ ODE used by SD3.5 [10, §4.3], the convention we adopt.

4

Concluding Remarks and Future Work

This thesis studied multi-entity spatial-layout control in text-to-image models built on multi-modal diffusion transformers (MM-DiT), where backbones such as SD3.5 and FLUX render high-fidelity images but place and separate entities unreliably once a prompt specifies several spatial relations at once. We introduced FOCAL, a training-free, plug-and-play controller that casts bounding-box-conditioned layout guidance as a stochastic-optimal-control (SOC) problem on the rectified-flow sampler and realises it as a single-pass velocity correction $\tilde{v}_\theta = v_\theta - w(t)\,\nabla_X f$ read from the network's own joint-attention. The controller leaves the pretrained weights untouched, adds one forward-backward pass per denoising step, and acts only on the sampling velocity, so the same algorithm transfers across backbones without retraining.

4.1. Summary of contributions

The thesis develops three contributions. First, FOCAL is, to our knowledge, the first explicit SOC formulation of bounding-box layout guidance on MM-DiT; rather than solving an inner optimization at each step, we correct the sampler's velocity once per step with a closed-form update derived from the unified objective. Second, we unify attention disentanglement and bounding-box conditioning in a single running cost $f = w_d L_d + w_t L_t + w_c L_c$, where L_d separates the per-entity attention maps, L_t pulls each entity's attention centroid toward its target box center, and L_c contains the remaining mass inside the box. The unification rests on the hypothesis that placement and separation are one coupled problem, since an entity cannot be moved cleanly into its box while its attention still overlaps another's, so the two are optimized jointly at every step rather than in sequence. Third, we curate SGE-mini, a 1,000-prompt benchmark drawn from SpatialGenEval [32] and restricted to the five spatial categories a region-based controller can affect, namely Object, Position, Orientation, Proximity, and Occlusion, so that the evaluation isolates multi-entity placement from properties no attention-region method can move.

Applied unchanged to SD3.5-Medium and FLUX.1-dev, FOCAL improves every SGE-mini category and every PosEval task: on SGE-mini it raises the SD3.5-Medium average accuracy by 0.069 and its spatial accuracy by 0.078, with the largest gain on Position (+0.140), the axis the translation and containment terms act on most directly, and the same controller lifts FLUX.1-dev by 0.060 and 0.062 respectively. On PosEval it raises average Soft-TIFA from 0.31 to 0.82 on SD3.5-Medium and to 0.80 on FLUX.1-dev, with the margin over the baselines widening as objects and relations accumulate, and both guided backbones attain the two highest average Soft-TIFA scores of any compared method, training-free or training-based alike, while SD3.5-Medium leads on average VQAScore (0.88), ahead of the far larger HiDream-O1-Image [6] (0.87), without any training of its own.

4.2. Limitations

The controller acts only on where attention mass lands on the image plane, so properties orthogonal to planar placement are largely untouched; Orientation barely improves under guidance (+0.008 on

SD3.5-Medium), because the way an entity faces is not a function of the position of its attention centroid or of how well that mass is contained, and the running cost has no term that targets it.

Similarly, Proximity and Occlusion are inherently three-dimensional, yet the running cost reflects only two dimensions. These categories do improve substantially (+0.097 and +0.065 on SD3.5-Medium), which we attribute to the frozen model's latent geometry rendering the occlusion ordering once the planar layout removes the ambiguity [9], rather than to any explicit depth reasoning in the guidance approach, so relations that require a depth ordering not implied by the planar arrangement remain outside its reach.

Placement is also limited by the two-stage LLM pipeline that produces the target layouts, since the guidance of FOCAL can only steer toward the layout it is given. Thus, a poorly parsed scene graph or an implausible layout caps the achievable alignment regardless of how well the velocity correction performs.

Finally, the global guidance strength λ does not transfer across architectures and must be set per backbone (500 on FLUX.1-dev, 1500 on SD3.5-Medium); the cost weights w_d, w_t, w_c transfer unchanged, but pushing λ beyond its calibrated value improves placement no further and visibly degrades image fidelity.

4.3. Broader Reflection

FOCAL is a response to an architectural shift in the field. The training-free layout methods that preceded it relied on the one-way text-to-image cross-attention of UNet models, which the joint attention of MM-DiT backbones dissolves, so spatial control had to be rebuilt on the new architecture. FOCAL does this by joining the stochastic-optimal-control view of guidance, the current state-of-the-art guidance approach, developed by Adjoint Matching [10] and FOCUS [4], to a bounding-box objective, and it keeps the weights of its chosen backbone diffusion model frozen. Compared against the field's other main direction, the scaling of generative T2I models such as Qwen-Image [33] and HiDream-O1-Image [6], the result here shows that a guided 2.5B model reaches placement accuracy competitive with far larger ones, which positions training-free control as a cheaper route to spatial precision. This leaves open whether future base models will become spatially reliable enough to need no external guidance at all, or whether pairing a strong frozen model with a lightweight correction will stay cheaper than depending on scale alone. The work also reflects the field's turn toward decomposed, VLM-judged spatial evaluation [32], to which SGE-mini contributes a benchmark that isolates the relations a region-based method can actually affect.

4.4. Applicability of FOCAL

Because FOCAL adds guidance to a frozen model rather than new generative capability, its applications and its stakeholders follow from that property. It can provide designers and non-expert users with layout-faithful generation that can be specified in plain language while costing model providers no retraining, since the same guidance transfers across backbones. The same precision carries some intrinsic risks: it lowers the effort needed to compose convincing fabricated scenes, and because the layout is produced by an LLM, its spatial and cultural priors are imported into every guided generation. These risks argue for responsible deployment with bias auditing of the layout model. Additionally, watermarking generated images with this guidance can prevent scenes fabricated through our method from passing as authentic, by giving downstream detectors and viewers a reliable signal of synthetic origin.

4.5. Future directions

The most direct extension follows from the limitations above: adding running-cost terms that target the axes the present cost ignores, so that an orientation term operating on the angular structure of an entity's attention, or a depth-aware containment that scores occlusion ordering rather than only planar position, would let the controller address Orientation and the three-dimensional relations it currently improves only as a by-product.

FOCAL sets the SOC terminal cost $g \equiv 0$ and encodes the entire objective in the running cost, but

the framework also admits a non-zero g on the final state, so adding an image-level reward such as a preference score [16] would let the controller trade spatial precision against image quality directly rather than only through the strength λ.

Because the guidance is stated without reference to any specific MM-DiT, extending it to further backbones like Qwen-Image, which appears here only as a quality ceiling, is a natural next target, and applying FOCAL to it would test whether the gains observed on SD3.5 and FLUX hold on a larger, stronger base model, while reducing the per-backbone calibration of λ to a single transferable or auto-tuned strength would remove the one remaining piece of manual configuration in an otherwise architecture-agnostic method.

Acknowledgement of AI Tools

In the course of this thesis, I used Anthropic's Claude as a support tool while remaining fully in the driver's seat of the work. The research direction, the design of FOCAL, its implementation, the experiments, and the interpretation of all results are my own. I used Claude in an assisting capacity: to improve the clarity, phrasing, and structure of the writing; to format and debug LaTeX for tables, figures, and the bibliography; to help locate and cross-check related work, which I then read and verified against the original papers; and as a useful tool to debug code. I treated its output as a draft to be checked rather than an authority: every claim, number, citation, and figure in this thesis was verified by me against primary sources and my own experimental records. I take full responsibility for the entire content of this work, including any remaining errors.

References

[1] Michael S. Albergo and Eric Vanden-Eijnden. "Building Normalizing Flows with Stochastic Interpolants". In: *International Conference on Learning Representations (ICLR)*. 2023. URL: https://openreview.net/forum?id=li7qeBbCR1t.

[2] Jessica Bader et al. *Stitch: Training-Free Position Control in Multimodal Diffusion Transformers*. 2025. arXiv: 2509.26644 [cs.CV].

[3] Julius Berner, Lorenz Richter, and Karen Ullrich. "An Optimal Control Perspective on Diffusion-Based Generative Modeling". In: *Transactions on Machine Learning Research (TMLR)* (2024). arXiv:2211.01364.

[4] Eric Tillmann Bill, Enis Simsar, and Thomas Hofmann. "FOCUS: Optimal Control for Multi-Entity World Modeling in Text-to-Image Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2026.

[5] Black Forest Labs. *FLUX.1*. https://github.com/black-forest-labs/flux. 2024.

[6] Qi Cai et al. *HiDream-O1-Image: A Natively Unified Image Generative Foundation Model with Pixel-level Unified Transformer*. 2026. arXiv: 2605.11061 [cs.CV].

[7] Hila Chefer et al. "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models". In: *ACM Transactions on Graphics (SIGGRAPH)* 42.4 (2023). arXiv: 2301.13826.

[8] Anthony Chen et al. "Training-Free Regional Prompting for Diffusion Transformers". In: *arXiv preprint arXiv:2411.02395* (2024). URL: https://arxiv.org/abs/2411.02395.

[9] Yida Chen, Fernanda Viégas, and Martin Wattenberg. "Beyond Surface Statistics: Scene Representations in a Latent Diffusion Model". In: *ICLR 2024 Workshop on Representational Alignment (Re-Align)*. 2024. arXiv: 2306.05720.

[10] Carles Domingo-Enrich et al. "Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryless Stochastic Optimal Control". In: *International Conference on Learning Representations (ICLR)*. arXiv:2409.08861. 2025.

[11] Patrick Esser et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2024.

[12] Wendell H. Fleming and Halil Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. 2nd. Springer, 2006.

[13] Aaron J. Havens et al. "Adjoint Sampling: Highly Scalable Diffusion Samplers via Adjoint Matching". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2025.

[14] Jonathan Ho and Tim Salimans. "Classifier-Free Diffusion Guidance". In: *arXiv preprint arXiv:2207.12598* (2022). URL: https://arxiv.org/abs/2207.12598.

[15] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*. 2014. URL: https://arxiv.org/abs/1312.6114.

[16] Yuval Kirstain et al. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[17] Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. "GrounDiT: Grounding Diffusion Transformers via Noisy Patch Transplantation". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[18] Yuheng Li et al. "GLIGEN: Open-Set Grounded Text-to-Image Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. arXiv: 2301.07093.

[19] Daniel Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012.

[20] Yaron Lipman et al. “Flow Matching for Generative Modeling”. In: *International Conference on Learning Representations (ICLR)*. 2023.

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. “Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow”. In: *International Conference on Learning Representations (ICLR)*. 2023.

[22] William Peebles and Saining Xie. “Scalable Diffusion Models with Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[23] Quynh Phung, Songwei Ge, and Jia-Bin Huang. “Grounded Text-to-Image Synthesis with Attention Refocusing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[24] Dustin Podell et al. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”. In: *International Conference on Learning Representations (ICLR)*. 2024.

[25] L. S. Pontryagin et al. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, 1962.

[26] Alec Radford et al. “Learning Transferable Visual Models from Natural Language Supervision”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[27] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.

[28] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. 2nd ed. Springer Series in Synergetics. Springer, 1996.

[29] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[30] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations (ICLR)*. 2021.

[31] Luran Wang et al. “Training Free Guided Flow-Matching with Optimal Control”. In: *International Conference on Learning Representations (ICLR)*. arXiv:2410.18070. 2025.

[32] Zengbin Wang et al. “Everything in Its Place: Benchmarking Spatial Intelligence of Text-to-Image Models”. In: *International Conference on Learning Representations (ICLR)*. 2026.

[33] Chenfei Wu et al. *Qwen-Image Technical Report*. 2025. arXiv: 2508.02324 [cs.CV].

[34] Jiayu Xiao et al. “R&B: Region and Boundary Aware Zero-Shot Grounded Text-to-Image Generation”. In: *International Conference on Learning Representations (ICLR)*. 2024.

[35] Jinheng Xie et al. “BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[36] Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer, 1999.

[37] Hui Zhang et al. “CreatiLayout: Siamese Multimodal Diffusion Transformer for Creative Layout-to-Image Generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025.

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. arXiv: 2302.05543.

[39] Dewei Zhou et al. “DreamRenderer: Taming Multi-Instance Attribute Control in Large-Scale Text-to-Image Models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025.