

OPTIMAL CONTROL MEETS FLOW MATCHING: A PRINCIPLED ROUTE TO MULTI-SUBJECT FIDELITY

Anonymous authors

Paper under double-blind review



Figure 1: **Optimal control makes flow matching models reliable on multi-subject prompts.** Using FOCUS at test time or via fine-tuning yields faithful multi-subject compositions with correct attributes, minimal leakage, and no omissions, while preserving base style.

ABSTRACT

Text-to-image (T2I) models excel on single-entity prompts but struggle with multi-subject descriptions, often showing attribute leakage, identity entanglement, and subject omissions. We introduce the first theoretical framework with a principled, optimizable objective for steering sampling dynamics toward multi-subject fidelity. Viewing flow matching (FM) through stochastic optimal control (SOC), we formulate subject disentanglement as control over a trained FM sampler. This yields two architecture-agnostic algorithms: (i) a training-free test-time controller that perturbs the base velocity with a single-pass update, and (ii) Adjoint Matching, a lightweight fine-tuning rule that regresses a control network to a backward adjoint signal while preserving base-model capabilities. The same formulation unifies prior attention heuristics, extends to diffusion models via a flow–diffusion correspondence, and provides the first fine-tuning route explicitly designed for multi-subject fidelity. Empirically, on Stable Diffusion 3.5, FLUX, and Stable Diffusion XL, both algorithms consistently improve multi-subject alignment while maintaining base-model style. Test-time control runs efficiently on commodity GPUs, and fine-tuned controllers trained on limited prompts generalize to unseen ones. We further highlight FOCUS (Flow Optimal Control for Unentangled Subjects), which achieves state-of-the-art multi-subject fidelity across models.

1 INTRODUCTION

Text-to-image (T2I) generators have made substantial progress in visual fidelity and prompt adherence, yet they remain brittle on *multi-subject* prompts. Typical failure modes include attribute leakage (an attribute intended for one subject propagates to others), identity entanglement (multiple subjects merged into a hybrid), and subject omission (Chefer et al., 2023; Liu et al., 2022; Bar-Tal et al., 2023; Dahary et al., 2024). These limitations hinder downstream applications such as story illustration, multi-panel composition, and scientific communication, where preserving subject identity and attribute binding is essential.

A unifying theoretical perspective on modern T2I generators is *flow matching* (FM), which parameterizes generation as a time-dependent flow from a base distribution to the data distribution via a learned vector field (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023). This framework encompasses both rectified-flow (RF) models used in recent large-scale systems such as Stable Diffusion 3.5 (Esser et al., 2024), FLUX (Labs et al., 2025), and earlier denoising-diffusion architectures such as Stable Diffusion 1.5 (Rombach et al., 2022), Stable Diffusion XL (Podell et al., 2024), enabling statements that transfer across architectures and training choices. We leverage this common ground to analyze—and improve—multi-subject fidelity in FM models.

Prior work has attempted to mitigate entanglement through *test-time* heuristics that reshape cross-attention (Meral et al., 2024) or adjust guidance (Feng et al., 2023), including token amplification (Chefer et al., 2023), constraint-based binding (Li et al., 2023b), and structure-aware attention editing (Hertz et al., 2023; Dahary et al., 2024). While effective in specific settings, these methods are heuristic and lack a unifying optimization objective, making it unclear when and why they succeed. Furthermore, most were developed for Stable Diffusion 1.x backbones, and their portability to RF and modern FM models remains limited.

In this work, we show that multi-subject *disentanglement* can be formulated as a *stochastic optimal control* (SOC) problem for trained FM-based samplers. Concretely, augmenting the base dynamics with a small control that balances proximity to the original generator against a differentiable *disentanglement objective* yields a principled formulation and two complementary algorithms:

- (i) **Test-time controller.** A lightweight single-pass controller derived from the optimality conditions of the SOC objective that steers sampling toward disentangled renderings without re-training. The formulation accepts *any* differentiable cost, thereby providing a principled path to adapt existing heuristics to modern FM models.
- (ii) **Fine-tuning via Adjoint Matching.** A stable, low-cost update rule based on *Adjoint Matching* (Domingo-Enrich et al., 2025) that regresses a control network onto a backward adjoint signal under a *memoryless* noise schedule, directly minimizing the disentanglement objective while preserving the base model’s style and support.

Empirically, our methods improve multi-subject fidelity across both modern FM models (Stable Diffusion 3.5, FLUX) and earlier diffusion backbones (Stable Diffusion XL). The test-time controller provides consistent gains with negligible overhead, while fine-tuning further reduces entanglement without degrading style or generalization beyond the training prompts. Building on these insights, we introduce FOCUS (Flow Optimal Control for Unentangled Subjects), which consolidates our framework into a practical algorithm and achieves the strongest results in our experiments. To foster transparency and reproducibility, we will release code, the curated dataset, and checkpoints of the best-performing fine-tuned models.

2 PRELIMINARIES

Flow Matching (FM) (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023) trains a time-dependent vector field $v_\theta : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ that transports a base distribution π_0 (e.g., $\mathcal{N}(0, I)$) to a target distribution (e.g. P_{data}), without simulating a forward noising process during training. Given a *reference path* $\bar{\mathbf{X}} = (\bar{X}_t)_{t \in [0, 1]}$ with $\bar{X}_0 \sim \pi_0$ and $\bar{X}_1 \sim \pi_1$, FM regresses the *conditional velocity*

$$u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1) := \frac{d}{dt} \bar{X}_t \quad (1)$$

so that $v_\theta(x, t)$ matches its conditional expectation $\mathbb{E}[u_t \mid \bar{X}_t = x]$ (Lipman et al., 2023).

Reference paths. A standard choice is the linear (Gaussian) interpolant

$$\bar{X}_t = \beta_t \bar{X}_0 + \alpha_t \bar{X}_1, \quad \alpha_0 = 0, \beta_0 = 1, \alpha_1 = 1, \beta_1 = 0, \quad (2)$$

where $(\alpha_t, \beta_t)_{t \in [0,1]}$ is a differentiable scheduler with α_t strictly increasing, and β_t strictly decreasing. The pathwise derivative is then $u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1) = \dot{\beta}_t \bar{X}_0 + \dot{\alpha}_t \bar{X}_1$.¹ A widely used instance is *rectified flow* (RF) with $\alpha_t = t$ and $\beta_t = 1 - t$ (Liu et al., 2023).

Training objective. FM is trained with the *conditional flow matching* loss (Lipman et al., 2023)

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1]} \mathbb{E}_{\substack{\bar{X}_0 \sim \pi_0 \\ \bar{X}_1 \sim \pi_1}} \left[\|v_\theta(\bar{X}_t, t) - u_t(\bar{X}_t \mid \bar{X}_0, \bar{X}_1)\|_2^2 \right], \quad (3)$$

which regresses the pathwise velocity toward its conditional mean at uniformly sampled times.

Sampling. After training, sample $X_0 \sim \pi_0$ and evolve the learned flow by solving the ODE

$$dX_t = v_\theta(X_t, t) dt, \quad (4)$$

which produces a path $(X_t)_{t \in [0,1]}$ whose marginals match those of the reference path $(\bar{X}_t)_{t \in [0,1]}$ under standard existence–uniqueness conditions; in particular $X_1 \sim \pi_1$ Lipman et al. (2023). More generally, FM admits a stochastic formulation (Domingo-Enrich et al., 2025) in which the drift is augmented by a arbitrary schedule-dependent correction with diffusion coefficient $\sigma(t) \geq 0$:

$$dX_t = \underbrace{\left(v_\theta(X_t, t) + \frac{\sigma(t)^2}{2\beta_t \left(\frac{\dot{\alpha}_t}{\alpha_t} \beta_t - \dot{\beta}_t \right)} \left(v_\theta(X_t, t) - \frac{\dot{\alpha}_t}{\alpha_t} X_t \right) \right)}_{:=b(X_t, t)} dt + \sigma(t) dB_t, \quad (5)$$

where $(B_t)_{t \geq 0}$ is standard Brownian motion in $\mathbb{R}^d$. We will refer to $b(X_t, t)$ as the (base) *drift*. Setting $\sigma \equiv 0$ recovers the deterministic ODE.

Connection to denoising diffusion. Classical denoising diffusion models arise as special cases of FM when their discrete procedures are lifted to continuous time; refer to Appendix C for details.

3 METHODOLOGY

We formulate disentanglement as optimal control over flow-matching dynamics, derive single-pass test-time and fine-tuned controllers, and introduce a probabilistic attention loss, FOCUS.

3.1 STOCHASTIC OPTIMAL CONTROL

Our goal is to reduce multi-subject entanglement while remaining close to the base model. To this end, we introduce a small *control* $u : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ into the drift and pose generation as a quadratic, control-affine SOC problem:

$$\min_{u \in \mathcal{U}} \mathbb{E} \left[\int_0^1 \frac{1}{2} \|u(X_t^u, t)\|_2^2 + f(X_t^u, t) dt + g(X_1^u) \right], \quad (6)$$

$$\text{s.t. } dX_t^u = (b(X_t^u, t) + \sigma(t)u(X_t^u, t)) dt + \sigma(t)dB_t, \quad X_0^u \sim \pi_0, \quad (7)$$

where X_t^u is the latent state, b is the base FM drift, $\sigma(t) \geq 0$ is a scalar diffusion schedule, and $(B_t)_{t \in [0,1]}$ is Brownian motion. The running cost $f : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$ will measure subject entanglement (e.g. $f \equiv \text{FOCUS}$), and we set the terminal cost $g \equiv 0$ in all derivations and experiments.

For control-affine dynamics with $\ell(x, u, t) = \frac{1}{2} \|u\|_2^2 + f(x, t)$, the Hamiltonian of the SOC is

$$\mathcal{H}(x, u, a, t) = \frac{1}{2} \|u\|_2^2 + f(x, t) + a^\top (b(x, t) + \sigma(t)u), \quad (8)$$

¹Over-dot denotes the time derivative, i.e., $\dot{x}_t = \frac{d}{dt}x_t$.

where $a(t) \in \mathbb{R}^d$ is the co-state (adjoint). Since $\mathcal{H}$ is strictly convex in u , the first-order optimality condition yields

$$u_t^* = -\sigma(t)a(t), \quad (9)$$

with adjoint dynamics

$$\frac{d}{dt}a(t) = -[\nabla_X b(X_t^u, t)^\top a(t) + \nabla_X f(X_t^u, t)], \quad a(1) = \nabla_X g(X_1^u). \quad (10)$$

3.2 ON-THE-FLY DISENTANGLEMENT (TEST-TIME CONTROL)

At inference, we solve Equation (6) *per trajectory* with frozen model parameters. The idea is to compute u_t^* on-the-fly and steer the sampling process at each timestep t . Directly computing u_t^* requires the adjoint $a(t)$ in Equation (9), which is defined along the *controlled* path via Equation (10). This is impractical because $a(t)$ depends on the terminal condition $a(1) = \nabla_X g(X_1^u)$, which depends on the endpoint X_1^u , which in turn depends on the future segment $(X_\tau^u)_{\tau \in [t, 1]}$; coupling a backward solve to the forward pass at every step. To obtain a *single-pass* controller, we approximate $a(t)$ locally at the current state. Concretely, we linearize Equation (10) around X_t^u , freeze $\nabla_X b \approx 0$, and treat the future state as locally constant:

$$a(t) \approx \int_t^1 \nabla_X f(X_t^u, \tau) d\tau \approx (1-t) \nabla_X f(X_t^u, t), \quad (11)$$

where the last step uses a left-Riemann approximation. Substituting into Equation (9) yields the instantaneous control

$$u_t^* \approx -\sigma(t)(1-t) \nabla_X f(X_t^u, t). \quad (12)$$

The approximation $\nabla_X b \approx 0$ is common in online control settings (Havens et al., 2025).

Velocity reparameterization (SDE). Let v_{base} denote the base FM velocity. We adopt the *memory-less* diffusion schedule, which makes the stochastic interpolant endpoints independent ($X_0 \perp X_1$) and yields a simple drift-velocity identity:

$$\sigma_{\text{mem}}(t) = \sqrt{2\beta_t \left(\frac{\dot{\alpha}_t}{\alpha_t} \beta_t - \dot{\beta}_t \right)}. \quad (13)$$

Under this choice, the drift-velocity relation from Equation (5) simplifies to $b(X_t, t) = 2v_\theta(X_t, t) - \frac{\dot{\alpha}_t}{\alpha_t} X_t$. Adding $+\sigma_{\text{mem}}(t)u_t$ to the drift shifts the velocity by $+\frac{1}{2}\sigma_{\text{mem}}(t)u_t$. Therefore the controlled velocity is

$$v_t^* = v_{\text{base}}(X_t, t) + \frac{\sigma_{\text{mem}}(t)}{2} u_t^* \approx v_{\text{base}}(X_t, t) - \frac{\sigma_{\text{mem}}^2(t)}{2} (1-t) \nabla_X f(X_t, t), \quad (14)$$

which can be passed to any SDE solver without modifying the integrator.²

Deterministic alternative (ODE). Many off-the-shelf T2I models are optimized for ODE sampling ($\sigma \equiv 0$). Decoupling σ from the control gives

$$\min_u \mathbb{E} \left[\int_0^1 \frac{1}{2} \|u(X_t, t)\|_2^2 + f(X_t, t) dt \right] \quad (15)$$

$$\text{s.t. } dX_t = (v_{\text{base}}(X_t, t) + u(X_t, t)) dt, \quad X_0 \sim \pi_0. \quad (16)$$

The Hamiltonian $\mathcal{H} = \frac{1}{2} \|u\|^2 + f + a^\top (v_{\text{base}} + u)$ yields $u_t^* = -a(t)$, and with the same local approximation:

$$v_t^* = v_{\text{base}}(X_t, t) - a(t) \approx v_{\text{base}}(X_t, t) - (1-t) \nabla_X f(X_t, t), \quad (17)$$

²If desired, the factor $\frac{1}{2}\sigma_{\text{mem}}(t)^2$ can be absorbed into the weight of f , yielding a schedule-invariant update.

3.3 FINE-TUNING FOR DISENTANGLEMENT

Our goal is to learn a control network u_θ that remains close to the base dynamics and generalizes beyond the specific trajectories used during training.

Adjoint Matching (AM). Directly solving Equation (10) during training is prohibitive because the adjoint $a(t)$ depends on the controlled path X_t^u . Instead, we use *Adjoint Matching* (Domingo-Enrich et al., 2025), regressing u_θ to a cheaper *lean adjoint* $\tilde{a}$ computed along frozen forward trajectories $(X_t)_{t \in [0,1]}$ while dropping u -dependent Jacobian terms:

$$\frac{d}{dt}\tilde{a}(t) = -[\nabla_X b(X_t, t)^\top \tilde{a}(t) + \nabla_X f(X_t, t)], \quad \tilde{a}(1) = \nabla_X g(X_1). \quad (18)$$

Memoryless training. To ensure that the learned control generalizes beyond the specific trajectories used in training, we follow Domingo-Enrich et al. (2025) and train under a *memoryless* generative process where $X_0 \perp X_1$, i.e., $p(X_0, X_1) = p(X_0)p(X_1)$. For linear (Gaussian) FM paths with scheduler $(\alpha_t, \beta_t)_{t \in [0,1]}$, the diffusion coefficient σ_{mem} from Equation (13) achieves this independence and makes the regression target *trajectory-stationary*.

Training objective. Each iteration proceeds as follows: (i) sample forward trajectories $(X_t)_{t \in [0,1]}$ under σ_{mem} with the current model frozen via Equation (5); (ii) integrate the lean adjoint $(\tilde{a}(t))_{t \in [0,1]}$ backward with Equation (18); (iii) regress the control toward the stationary target $-\sigma_{\text{mem}}(t)\tilde{a}(t)$ by minimizing

$$\mathcal{L}_{\text{AM}}(\theta) := \frac{1}{2} \int_0^1 \|u_\theta(X_t, t) + \sigma_{\text{mem}}\tilde{a}(t)\|^2 dt. \quad (19)$$

The memoryless schedule is only required during *fine-tuning*. At inference $\sigma(t)$ can be set to zero, allowing to use faster off-the-shelf ODE samplers.

3.4 MEASURING MULTI-SUBJECT ENTANGLEMENT

At each sampling step, T2I backbones compute *cross-attention* from image-space queries to text tokens. Empirically, these token-wise cross-attention maps correlate with the eventual spatial placement of the corresponding entities (Hertz et al., 2023; Chefer et al., 2023). This enables us to diagnose and mitigate subject entanglement *during* generation by measuring spatial interactions among *subject-specific* attention maps, rather than relying solely on post-hoc image encoders (see Figure 2).

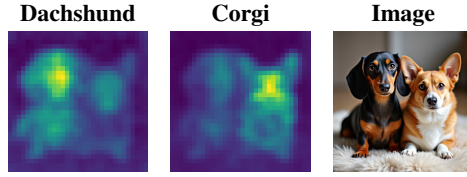


Figure 2: Extracted cross-attention maps for both subjects in FLUX.1 [dev].

Most prior work that optimizes multi-subject behavior via cross-attention treats these maps as generic similarity scores (e.g., maximizing cosine similarity (Meral et al., 2024) or activation differences (Chefer et al., 2023)). However, cross-attention arises from a softmax: each map is a *probability distribution* over spatial locations. Ignoring this structure discards a principled probabilistic footing and can induce artifacts such as over-concentration. We instead treat attention maps as distributions and optimize them accordingly.

FOCUS. Let d denote the number of spatial locations and let Δ^{d-1} be the probability simplex. For a finite set $P = \{vp_1, \dots, p_n\} \subset \Delta^{d-1}$ of distributions, define the Jensen–Shannon divergence

$$D_{\text{JS}}(P) = \frac{1}{n} \sum_{i=1}^n D_{\text{KL}}(p_i \| m), \quad m = \frac{1}{n} \sum_{j=1}^n p_j,$$

with $D_{\text{KL}}(p \| q) = \sum_{i=1}^d p_i \log \frac{p_i}{q_i}$ being the Kullback-Leibler divergence. Since $D_{\text{JS}}(P) \in [0, \log n]$, we normalize by dividing with $\log n$ to obtain $\hat{D}_{\text{JS}}(P) \in [0, 1]$, which makes scores comparable across different set sizes; see Theorem B.1 for a proof of this upper bound.

We introduce FOCUS (Flow Optimal Control for Unentangled Subjects) to encourage, for each subject, *unimodal, spatially localized, and nonoverlapping* attention. Let S be the set of subjects in



Figure 3: Qualitative results with test-time control on Stable Diffusion 3.5 and FLUX.1. Each heuristic is shown at its optimal λ . Additional examples appear in Figures 9 and 10 of the Appendix.

the prompt. For each subject $s \in S$, collect its attention maps at the current step into $P_s \subset \Delta^{d-1}$ (e.g., across layers or heads), and define the subject mean $\mathbf{m}_s = \frac{1}{|P_s|} \sum_{p \in P_s} p$. Let $M = \{\mathbf{m}_s \mid s \in S\}$ be the set of subject means. Our FOCUS loss combines *within-subject agreement* and *between-subject separation*:

$$\text{FOCUS}(S) = \frac{1}{2} \left(\frac{1}{|S|} \sum_{s \in S} \hat{D}_{JS}(P_s) \right) + \frac{1}{2} \left(1 - \hat{D}_{JS}(M) \right) \quad (20)$$

The first term penalizes dispersion within each subject’s maps (encouraging consistent binding, and for multi-encoder models such as SD 3.5, agreement across encoders). The second term rewards separation among subjects by maximizing divergence between their mean attention distributions. By construction, $\text{focus} \in [0, 1]$: 0 indicates perfect disentanglement (low intra-subject dispersion and maximal inter-subject separation), while larger values indicate greater entanglement.

4 RELATED WORK

We review approaches to *multi-subject* T2I generation. We first cover *training-free* attention-space interventions that operate at inference time. We then discuss methods that enforce *regional/layout* constraints or combine multiple diffusion paths. Finally, we survey *training-time* objectives that strengthen subject–attribute binding.

Attention-space interventions (training-free). A large body of work steers pre-trained generators at inference by manipulating cross-attention. At each sampling step, the model produces attention weights from spatial queries to text tokens; selecting the column for a token and normalizing over space yields a token-conditioned spatial map. Methods then *assess* entanglement (e.g., by measuring overlap across subjects) and *modify* attention or latents to promote coverage and separation.

Attend&Excite amplifies token activations to enforce entity coverage and reduce neglect or leakage (Chefer et al., 2023). *Divide&Bind* adds an inference-time objective that separately enforces subject coverage and attribute binding, optimizing latents during sampling (Li et al., 2023b). *Structured Diffusion Guidance* injects linguistic structure (e.g., dependency trees) to guide attention manipulation for multi-object composition (Feng et al., 2023). *Prompt-to-Prompt* locks cross-attention correspondences to preserve word–subject alignments across edits, often used to maintain multi-subject layouts (Hertz et al., 2023). *CONFORM* formulates a contrastive, InfoNCE-style objective that separate different subjects while pulling subject–attribute pairs together (Meral et al., 2024).

While effective in specific setups, these methods are heuristic and lack a unifying optimization principle; moreover, many were developed for Stable Diffusion 1.x backbones, limiting portability to modern flow-matching models. In contrast, our method derives a controller from a single SOC

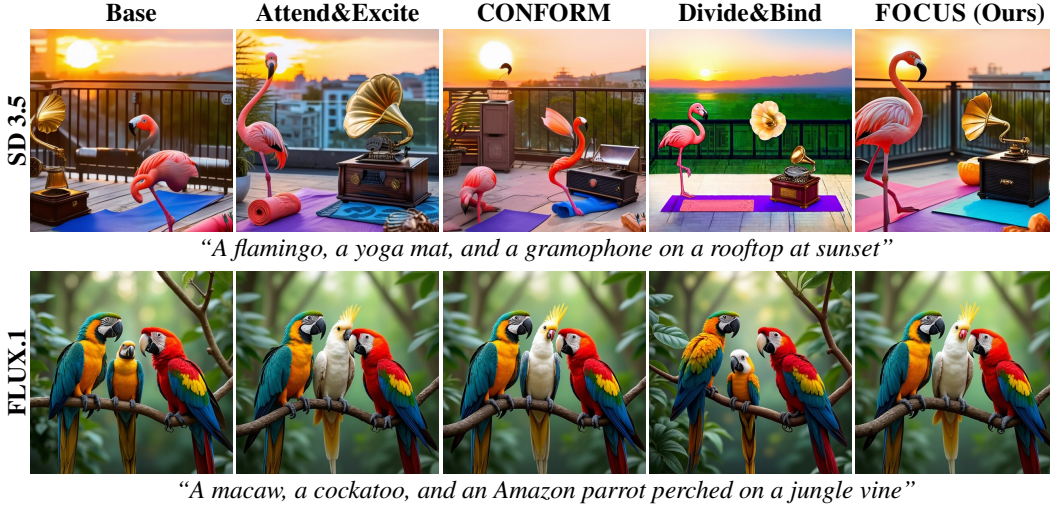


Figure 4: Qualitative results after fine-tuning Stable Diffusion 3.5 and FLUX.1. Each heuristic uses its optimal hyperparameters. Additional examples appear in Figures 11 and 12 of the Appendix.

objective at the FM level, yielding architecture-agnostic updates. We also instantiate our controller with costs derived from several of the above heuristics to demonstrate principled portability.

Regional/layout composition and multi-path fusion. A complementary direction constrains *where* subjects appear. *MultiDiffusion* fuses multiple diffusion trajectories under shared spatial constraints (e.g., boxes or masks), enabling faithful multi-subject placement without retraining (Bar-Tal et al., 2023). Related systems extend this idea to interactive, region-based workflows. *GLIGEN* augments a frozen backbone with grounding layers and conditions on bounding boxes or phrases to place multiple objects precisely (Li et al., 2023a). More recently, *Be Decisive* leverages the layout implicitly encoded in the initial noise and refines it during denoising, avoiding conflicts with externally imposed layouts and improving prompt alignment while preserving model priors (Dahary et al., 2025). These approaches disentangle primarily via spatial decoupling but often require user-specified or learned layouts, which increases user effort and restricts spontaneous subject interaction. Our method reduces entanglement without explicit spatial annotations, requiring only the text prompt (and its subjects).

Training-time objectives for multi-subject fidelity. Some works alter training signals to strengthen subject-attribute binding. *TokenCompose* introduces token-level supervision to improve consistency for prompts with multiple categories and attributes (Wang et al., 2024b). Region-aware objectives decompose complex prompts into per-region descriptions and enforce alignment, reducing cross-entity leakage. Such methods typically assume curated supervision and substantial retraining. In contrast, our fine-tuning objective is lightweight: it adapts pre-trained models via Adjoint Matching and requires only text prompts, while our test-time controller operates with zero parameter updates.

5 EXPERIMENTS

We evaluate our approach in three stages. We first describe datasets, metrics, models, and baselines. We then present *test-time* (on-the-fly) results, followed by *fine-tuning* results. All experiments ran on NVIDIA A100/H100 GPUs. While the test-time controller runs on commodity GPUs with as little as 12 GB VRAM, fine-tuning experiments fit within the VRAM of H100 GPUs.

Base Models. We report results on two open-source flow-matching models: *Stable Diffusion 3.5* (SD 3.5) (Esser et al., 2024) and *FLUX.1 [dev]* (FLUX.1) (Labs et al., 2025).

Dataset. We create a 150-prompt corpus with 2–4 subjects per prompt using GPT-5. Half the prompts contain *similar* subjects (e.g., “a black bear and a brown bear[...]”); the rest contain *dis-similar* subjects (e.g., “a snowboard, a telescope, and a husky[...]”). For each prompt, we annotate subject token indices for both CLIP and T5 encoders to extract cross-attention maps for the heuristics. Such per-subject annotations are typically absent from existing corpora.

Table 1: Test-time control results at the optimal λ for each heuristic. We report mean $\pm$ std over all prompts and seeds; the top three per metric are highlighted (gold/silver/bronze).

	Heuristic	CLIP I-T $\uparrow$	SigLIP-2 I-T $\uparrow$	BLIP T-T $\uparrow$	Qwen2 T-T $\uparrow$	PickScore I-T $\uparrow$	ImgRew I-T $\uparrow$	Composite $\uparrow$
SD 3.5	Base	0.3474 $\pm$ 0.03	0.2309 $\pm$ 0.05	0.5731 $\pm$ 0.15	0.6402 $\pm$ 0.08	22.6940 $\pm$ 0.99	1.3175 $\pm$ 0.68	0.0000 $\pm$ 0.00
	Attend&Excite	0.3484 $\pm$ 0.03	0.2326 $\pm$ 0.05	0.5752 $\pm$ 0.15	0.6404 $\pm$ 0.08	22.6950 $\pm$ 1.01	1.3545 $\pm$ 0.66	3.1714 $\pm$ 0.53
	CONFORM	0.3481 $\pm$ 0.03	0.2323 $\pm$ 0.05	0.5773$\pm$0.15	0.6421$\pm$0.08	22.7188 $\pm$ 0.99	1.3684 $\pm$ 0.64	3.4336 $\pm$ 0.45
	Divide&Bind	0.3489$\pm$0.03	0.2316 $\pm$ 0.05	0.5742 $\pm$ 0.14	0.6399 $\pm$ 0.08	22.6779 $\pm$ 1.03	1.3493 $\pm$ 0.67	3.9373 $\pm$ 0.78
	FOCUS (Ours)	0.3483 $\pm$ 0.03	0.2344$\pm$0.04	0.5751 $\pm$ 0.15	0.6385 $\pm$ 0.08	22.7499$\pm$1.02	1.4003$\pm$0.62	4.2865$\pm$0.86
FLUX.1	Base	0.3449 $\pm$ 0.03	0.2271 $\pm$ 0.05	0.5739 $\pm$ 0.15	0.6300 $\pm$ 0.09	23.4234 $\pm$ 1.03	1.2970$\pm$0.66	0.0000 $\pm$ 0.00
	Attend&Excite	0.3430 $\pm$ 0.03	0.2242 $\pm$ 0.05	0.5716 $\pm$ 0.14	0.6304 $\pm$ 0.09	23.2549 $\pm$ 1.11	1.2494 $\pm$ 0.70	1.7595 $\pm$ 0.67
	CONFORM	0.3436 $\pm$ 0.03	0.2252 $\pm$ 0.05	0.5726 $\pm$ 0.15	0.6321 $\pm$ 0.09	23.3574 $\pm$ 1.03	1.2461 $\pm$ 0.70	1.5114 $\pm$ 0.26
	Divide&Bind	0.3453$\pm$0.03	0.2272$\pm$0.05	0.5722 $\pm$ 0.15	0.6330$\pm$0.08	23.4395$\pm$1.02	1.2939 $\pm$ 0.67	1.6352 $\pm$ 0.44
	FOCUS (Ours)	0.3446 $\pm$ 0.03	0.2268 $\pm$ 0.05	0.5741$\pm$0.14	0.6326 $\pm$ 0.08	23.4274 $\pm$ 1.02	1.2913 $\pm$ 0.67	1.9712$\pm$0.31

Metrics. To quantify multi-subject fidelity, we follow Yu & Chien (2025) and report two alignment groups. For image–text (I–T) alignment we compute CLIP (Radford et al., 2018) and SigLIP-2 (Tschannen et al., 2025) cosine similarities between image and prompt embeddings. For caption-based text–text (T–T) faithfulness, we caption each image with BLIP (Li et al., 2022) and Qwen2-VL (Wang et al., 2024a) and measure semantic similarity to the prompt. We additionally report preference-trained scores, PickScore (Kirstain et al., 2023) and ImageReward (Xu et al., 2023), as proxies for human preference.

For model selection, we compute a composite score per hyperparameter combination by *macro-averaging* baseline-relative gains across metrics; see Appendix D.2 for the formula. Because we aim to preserve base style and subject depiction, global alignment scores may shift modestly even when multi-subject fidelity improves. Unless noted otherwise, we generate five images per prompt (distinct seeds) per hyperparameter setting, fixing sampler, steps, guidance, and resolution allowing to make direct comparisons between test-time control and fine-tuning. Full details are in Appendix D.

Baselines and heuristics. To demonstrate portability across FM models and legacy U-Net heuristics, we evaluate *Attend&Excite* (Chefer et al., 2023), *CONFORM* (Meral et al., 2024), *Divide&Bind* (Li et al., 2023b), and our heuristic FOCUS. Because cost magnitudes differ, we optimize a scaled running cost $\lambda \cdot f(X_t, t)$ with $\lambda > 0$. The effect of λ at test time is shown in Figure 7.

Human study. Automated metrics struggle to detect attribute leakage reliably (Dahary et al., 2025), so we conducted a prompt-conditioned, pairwise preference study with 50 participants. In each trial, participants viewed two images from our evaluation suite alongside the prompt and selected the image that better matched the prompt, yielding 2,000 pairwise judgments. From these outcomes we computed Elo ratings (across-method comparability) and win rates (fraction of pairwise wins).

5.1 ON-THE-FLY DISENTANGLEMENT (TEST-TIME CONTROL)

We sweep ten λ values per heuristic and select the best via the composite score defined above. Table 1 report per-heuristic, per-model results at the optimal λ , qualitative examples are shown in Figures 9 and 10, and Human Study results are summarized in Table 2.

All heuristics outperform the base sampler on SD 3.5 and FLUX.1, indicating that the SOC formulation yields a principled route to port legacy heuristics to modern FM models. Qualitatively, outputs show higher multi-subject fidelity: subjects are more often present and better separated than in the base model. Our human study shows similar trends, with higher win-rates and Elo ratings. While FOCUS is not best on every metric, it attains the highest composite score across all heuristics and achieves almost all best scores in our human study.

Table 2: Human preference study for test-time control. Report pairwise win rate and Elo rating; see appendix E for details.

Heuristic	SD3.5		FLUX.1	
	Win[%]	Elo $\uparrow$	Win[%]	Elo $\uparrow$
Base	45%	1517	46%	1464
Attend&Excite	53%	1500	49%	1526
CONFORM	42%	1373	50%	1498
Divide&Bind	50%	1562	50%	1450
FOCUS (Ours)	58%	1548	54%	1562

5.2 FINE-TUNING FOR DISENTANGLEMENT

We insert LoRA layers (Hu et al., 2022) into self-attention blocks and freeze all base parameters. We use rank $r=4$ (training $< 0.1\%$ of parameters). We sweep λ and other hyperparameters, including

Table 4: Fine-tuning results at the set of hyperparameters for each heuristic. We report mean $\pm$ std over all prompts and seeds; the top three per metric are highlighted (gold/silver/bronze).

Heuristic	CLIP I-T $\uparrow$	SigLIP-2 I-T $\uparrow$	BLIP T-T $\uparrow$	Qwen2 T-T $\uparrow$	PickScore I-T $\uparrow$	ImgRew I-T $\uparrow$	Composite $\uparrow$	
SD 3.5	Base	0.3474 $\pm$ 0.03	0.2309 $\pm$ 0.05	0.5731 $\pm$ 0.15	0.6402 $\pm$ 0.08	22.6940 $\pm$ 0.99	1.3175 $\pm$ 0.68	0.0000 $\pm$ 0.00
	Attend&Excite	0.3469 $\pm$ 0.03	0.2281 $\pm$ 0.04	0.5747 $\pm$ 0.15	0.6425 $\pm$ 0.08	22.8429 $\pm$ 1.01	1.4460 $\pm$ 0.60	5.7181 $\pm$ 1.21
	CONFORM	0.3478 $\pm$ 0.03	0.2294 $\pm$ 0.05	0.5646 $\pm$ 0.15	0.6393 $\pm$ 0.09	22.5962 $\pm$ 0.99	1.3782 $\pm$ 0.63	3.4583 $\pm$ 1.05
	Divide&Bind	0.3486 $\pm$ 0.03	0.2266 $\pm$ 0.05	0.5870 $\pm$ 0.14	0.6358 $\pm$ 0.08	22.3401 $\pm$ 0.99	1.3524 $\pm$ 0.68	0.8006 $\pm$ 0.69
	FOCUS (Ours)	0.3495 $\pm$ 0.03	0.2331 $\pm$ 0.04	0.5744 $\pm$ 0.15	0.6383 $\pm$ 0.08	22.6445 $\pm$ 0.97	1.4495 $\pm$ 0.58	5.9174 $\pm$ 1.19
FLUX.1	Base	0.3449 $\pm$ 0.03	0.2271 $\pm$ 0.05	0.5739 $\pm$ 0.15	0.6300 $\pm$ 0.09	23.4234 $\pm$ 1.03	1.2970 $\pm$ 0.66	0.0000 $\pm$ 0.00
	Attend&Excite	0.3468 $\pm$ 0.03	0.2320 $\pm$ 0.05	0.5876 $\pm$ 0.15	0.6382 $\pm$ 0.08	23.3333 $\pm$ 1.01	1.3806 $\pm$ 0.62	2.3477 $\pm$ 0.79
	CONFORM	0.3458 $\pm$ 0.03	0.2305 $\pm$ 0.04	0.5800 $\pm$ 0.15	0.6369 $\pm$ 0.08	23.3724 $\pm$ 1.00	1.3631 $\pm$ 0.63	1.9591 $\pm$ 0.83
	Divide&Bind	0.3445 $\pm$ 0.03	0.2296 $\pm$ 0.05	0.5705 $\pm$ 0.15	0.6246 $\pm$ 0.09	23.1909 $\pm$ 1.06	1.2269 $\pm$ 0.70	0.2002 $\pm$ 0.47
	FOCUS (Ours)	0.3468 $\pm$ 0.03	0.2328 $\pm$ 0.05	0.5780 $\pm$ 0.15	0.6386 $\pm$ 0.08	23.3278 $\pm$ 1.01	1.3899 $\pm$ 0.61	2.5881 $\pm$ 0.79

dataset choice; see Appendix D. With the best settings, fine-tuning takes 17 min on SD 3.5 and 79 min on FLUX.1. Table 4 reports metrics across all heuristics and models; qualitative results appear in Figures 4, 11 and 12; human preferences are summarized in Table 3.

Training data comprise two small subsets of our evaluation dataset. The first is a single prompt, “A horse and a bear in a forest,” where SD 3.5 fails reliably (HORSE&BEAR). The second contains 15 prompts with two semantically similar subjects (TWOOBJECTS). Despite their size, both subsets yield gains on diverse *unseen* prompts, including different subject categories, prompts with more than two subjects, and different subject token positions, suggesting that our method targets a core failure mode in multi-subject composition.

Across the board, the fine-tuned models outperform their test-time controlled counterparts. This matches our theory, since during fine-tuning the adjoint signal is computed explicitly over the full trajectory, whereas test-time control relies on a single-pass approximation. Between the two training sets, HORSE&BEAR yields the strongest gains with an 85% relative improvement in the composite scores in contrast to TWOOBJECTS for SD 3.5 and about 5% for FLUX.1. Across heuristics, FOCUS attains the highest composite score, indicating the largest average improvement across metrics. Consistently, in the human study FOCUS achieves the highest win rates against competing heuristics and among the highest Elo ratings for both models.

Table 3: Human preference study for fine-tuned models. Report pairwise win rate and Elo rating; see appendix E for details.

Heuristic	SD3.5		FLUX.1	
	Win[%]	Elo $\uparrow$	Win[%]	Elo $\uparrow$
Base	39%	1355	51%	1462
Attend&Excite	56%	1584	50%	1476
CONFORM	49%	1520	50%	1620
Divide&Bind	48%	1436	43%	1442
FOCUS (Ours)	57%	1605	54%	1500

5.3 CLASSICAL DENOISING DIFFUSION

Although our algorithms are derived for flow matching, Appendix C establishes a correspondence denoising diffusion. To test the theory, we apply the *test-time controller* to *Stable Diffusion XL*, a U-Net based denoising diffusion model. As shown in Figure 5, FOCUS improves on the prompt ‘A lion and a tiger resting side by side[...]’ by reducing attribute leakage.



Figure 5: Transfer to SDXL.

6 DISCUSSION AND FUTURE WORK

We propose a control-theoretic framework for multi-subject fidelity, instantiated either as a single-pass test-time controller or as a lightweight fine-tuned controller. The formulation accommodates existing attention-based heuristics, and our FOCUS yields the most consistent gains across settings. The two realizations offer complementary trade-offs: test-time control applies directly to a frozen model given subject tokens, at the cost of roughly $2\times$ longer inference, whereas fine-tuning requires subject tokens only during training and matches the base model’s inference speed during inference. Finally, the strong generalization of fine-tuning—even from a single prompt—suggests an underlying attention-level failure mode in current T2I models; future work should probe this mechanism and develop annotation-free proxies and automated subject tokenization.

REFERENCES

- Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: [https://doi.org/10.1016/0304-4149\(82\)90051-5](https://doi.org/10.1016/0304-4149(82)90051-5). URL <https://www.sciencedirect.com/science/article/pii/0304414982900515>.
- Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/bar-tal23a.html>.
- 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 (TOG)*, 42(4):1–10, 2023.
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pp. 432–448. Springer, 2024.
- Omer Dahary, Yehonathan Cohen, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be decisive: Noise-induced layouts for multi-subject generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.
- Carles Domingo-Enrich, Michal Drozdal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=xQBRrtQM8u>.
- 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, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/esser24a.html>.
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=PUIqjT4rzq7>.
- Aaron J Havens, Benjamin Kurt Miller, Bing Yan, Carles Domingo-Enrich, Anuroop Sriram, Daniel S. Levine, Brandon M Wood, Bin Hu, Brandon Amos, Brian Karrer, Xiang Fu, Guan-Hong Liu, and Ricky T. Q. Chen. Adjoint sampling: Highly scalable diffusion samplers via adjoint matching. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=6Eg1OrHmg2>.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.
- Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, June 2024.
- 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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL <https://arxiv.org/abs/2506.15742>.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/li22n.html>.
- 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)*, pp. 22511–22521, June 2023a.
- Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, pp. 366, 2023b. URL <http://proceedings.bmvc2023.org/366/>.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=PqvMRDCJT9t>.
- Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 423–439, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19790-1.
- Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=XVjTT1nw5z>.

- Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9005–9014, June 2024.
- Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72983–73007. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e6d58fc68c0f3c36ae6e0e64478a69c0-Paper-Conference.pdf.
- Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/nichol21a.html>.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=di52zR8xgf>.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1801.06146*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- 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)*, pp. 10684–10695, June 2022.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/sohl-dickstein15.html>.
- 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*, 2021. URL <https://openreview.net/forum?id=PXTIG12RRHS>.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL <https://arxiv.org/abs/2502.14786>.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024a. URL <https://doi.org/10.48550/arXiv.2409.12191>.
- Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8553–8564, June 2024b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

Hsiang-Chun Yu and Jen-Tsung Chien. Attention disentanglement for semantic diffusion modeling in text-to-image generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888688.

APPENDIX CONTENTS

A	Use of Large Language Models	15
B	FOCUS	15
C	Denoising Diffusion as Flow Matching	16
C.1	VP chain to SDE	16
C.2	Reverse-time dynamics	17
C.3	Time change to FM	17
C.4	Score relations	17
D	Hyperparameters	18
D.1	Sampling Parameters	18
D.2	Metrics	18
D.3	Test-Time Control	18
D.4	Fine-tuning	18
D.5	Additional Metric: Open-Vocabulary Detection	19
E	Human Study	20
E.1	Elo Rating Computation	20
F	Extra Samples	21
F.1	Test-Time Control: Stable Diffusion 3.5	21
F.2	Test-Time Control: FLUX.1 [dev]	22
F.3	Fine-tuned: Stable Diffusion 3.5	23
F.4	Fine-tuned: FLUX.1 [dev]	24

A USE OF LARGE LANGUAGE MODELS

We used a large language model (OpenAI GPT-5 via ChatGPT) for two purposes: (i) expanding a small, human-written set of text prompts to create additional prompts for our synthetic dataset, and (ii) polishing writing (grammar, clarity, and tone). For dataset construction, the model generated semantically similar prompt variants; all outputs were screened and curated by the authors. For writing, the authors drafted all sections and used the model only for copy-editing, not for introducing technical content.

B FOCUS

This appendix details FOCUS, our probabilistic attention heuristic used as a running cost for disentanglement. We emphasize three practical design choices: (i) encoding spatial proximity before measuring divergence, (ii) aggregating attention maps prior to scoring, and (iii) omitting an explicit collapse regularizer.

Spatially aware divergence. We promote separation of subjects by maximizing a Jensen–Shannon divergence (JSD) defined over attention distributions. A naïve computation on flattened maps discards locality, allowing distant activations to interact as if adjacent. To preserve spatial structure, we (i) reshape token-embedding maps to the target aspect ratio, (ii) apply a light 2D Gaussian smoothing, and only then (iii) flatten for scoring. This encodes proximity and mitigates grid-like artifacts during optimization.

Block selection and aggregation. Modern T2I backbones follow Diffusion Transformer designs (Peebles & Xie, 2023). Rather than computing scores *per block* and averaging their scores which can result in conflicting update directions, we first aggregate attention and then score. Concretely, we average cross-attention maps over all blocks that process text and image tokens *separately*, producing a single map per token and a consistent optimization direction. Blocks that jointly process text and image tokens are excluded from this aggregation for compatibility.

No explicit collapse regularizer. We experimented with an entropy-based regularizer aimed at discouraging overly concentrated (collapsed) attention. Let $H(\mathbf{p}) = -\sum_i \mathbf{p}_i \log \mathbf{p}_i$ denote the Shannon Entropy and $\hat{H}(\mathbf{p}) = H(\mathbf{p}) / \log d \in [0, 1]$ its normalized version, where d is the number of spatial locations. For each subject we form its mixture distribution $\mathbf{m}_s$ and added

$$\gamma_{\text{reg}} \cdot \frac{1}{|S|} \sum_{s \in S} \left(1 - \hat{H}(\mathbf{m})\right), \quad (21)$$

scaling by $\gamma_{\text{reg}} > 0$ to control its effect. In our experiments, small γ_{reg} made the term largely inactive, while larger γ_{reg} pushed mass away from the subject rather than stabilizing it, see Figure 6 for an example. We therefore omit this term in FOCUS and rely on the probabilistic objective described above.

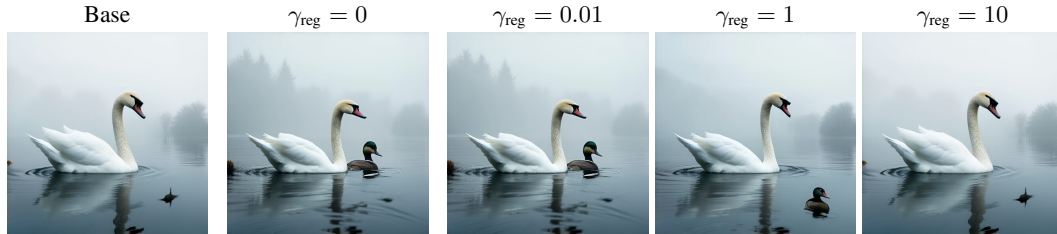


Figure 6: Ablation of regularizer strength γ_{reg} for the test-time controller on Stable Diffusion 3.5.

Lemma B.1 (Upper Bound of Jensen–Shannon Divergence). *Let $P = \{\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(n)}\} \subset \Delta^{d-1}$ be a set of probability distributions. Then, $D_{\text{JS}}(P)$ is upper bounded by $\log n$.*

Proof. Define P as in theorem B.1, then the JSD is defined as follows:

$$D_{\text{JS}}(P) = \frac{1}{n} \sum_{k=1}^n D_{\text{KL}}(\mathbf{p}^{(k)} \parallel \mathbf{m}), \quad \mathbf{m} = \frac{1}{n} \sum_{k=1}^n \mathbf{p}^{(k)}.$$

We can upper bound each D_{KL} -term as follows:

$$\begin{aligned} D_{\text{KL}}(\mathbf{p}^{(k)} \parallel \mathbf{m}) &= \sum_{i=1}^d p_i^{(k)} \log \frac{p_i^{(k)}}{m_i} \\ &= \sum_{i=1}^d p_i^{(k)} \log \frac{p_i^{(k)}}{\frac{1}{n} \sum_{\ell=1}^n p_i^{(\ell)}} \\ &= \sum_{i=1}^d p_i^{(k)} \log \left(n \cdot \frac{p_i^{(k)}}{\sum_{\ell=1}^n p_i^{(\ell)}} \right) \\ &\leq \sum_{i=1}^d p_i^{(k)} \log n \\ &= \log n. \end{aligned}$$

Plugging this bound back into the definition of the JSD, yields the desired results:

$$\frac{1}{n} \sum_{k=1}^n D_{\text{KL}}(\mathbf{p}^{(k)} \parallel \mathbf{m}) \leq \frac{1}{n} \sum_{k=1}^n \log n = \log n$$

□

Normalization. Because $D_{\text{JS}}(P) \in [0, \log n]$, we use the normalized score $\hat{D}_{\text{JS}}(P) = D_{\text{JS}}(P) / \log n \in [0, 1]$, which makes values comparable across different set sizes.

C DENOISING DIFFUSION AS FLOW MATCHING

This section makes precise how classical denoising diffusion (score-based) models arise as a special case of the flow-matching (FM) framework. We first derive the continuous-time SDE limit of the variance-preserving (VP) family (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021); after we express reverse-time generation; and finally show an explicit parameterization that uses a diffusion model as an FM velocity field. Analogous statements hold for VE and EDM variants (Nichol & Dhariwal, 2021; Karras et al., 2022; 2024).

C.1 VP CHAIN TO SDE

Let $X_0 \sim p_{\text{data}}$. The standard K -step VP forward noising chain is

$$X_k = \sqrt{\alpha_k} X_{k-1} + \sqrt{1 - \alpha_k} \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \mathbf{I}), \quad k = 1, \dots, K, \quad (22)$$

where $\alpha_k := 1 - \beta_k \in (0, 1)$ with $\beta_k \in (0, 1)$ typically increasing over k (Ho et al., 2020). This yields the closed-form marginal

$$X_k \mid X_0 \sim \mathcal{N}(\sqrt{\bar{\alpha}_k} X_0, (1 - \bar{\alpha}_k) \mathbf{I}), \quad \bar{\alpha}_k := \prod_{i=1}^k \alpha_i. \quad (23)$$

For sufficiently large K , X_K is approximately standard normal (Ho et al., 2020).

We lift this formulation to continuous time by defining a uniform grid $\tau_k := k/K$ on $[0, 1]$, so every increment is $\Delta\tau = 1/K$. Define a piecewise-constant rate $\beta(\tau)$ via $\beta(\tau) := \beta_k / \Delta\tau$ for $\tau \in [\tau_{k-1}, \tau_k)$. Then by using the first-order Taylor approximation of $\sqrt{1+x}$, we can rewrite

$\sqrt{\alpha_k} = \sqrt{1 - \beta_k} \approx 1 - \frac{1}{2}\beta_k + \mathcal{O}(\beta_k^2)$, and obtain

$$\Delta X_k := X_k - X_{k-1} \quad (24)$$

$$= -\frac{1}{2}\beta_k X_{k-1} + \sqrt{\beta_k}\epsilon_k + \mathcal{O}(\beta_k^2) \quad (25)$$

$$= \left(-\frac{1}{2}\beta(\tau_{k-1})X_{k-1}\right)\Delta\tau + \sqrt{\beta(\tau_{k-1})}\sqrt{\Delta\tau}\epsilon_k + \mathcal{O}\left((\Delta\tau)^{\frac{3}{2}}\right). \quad (26)$$

This is the Euler–Maruyama discretizations of the forward/diffusion VP-SDE:

$$dX_\tau = -\frac{1}{2}\beta(\tau)X_\tau d\tau + \sqrt{\beta(\tau)}dB_\tau, \quad \tau \in [0, 1], \quad (27)$$

and the discrete chain converges to this SDE as $K \rightarrow \infty$. Moreover, the SDE has Gaussian marginals

$$X_\tau | X_0 \sim \mathcal{N}\left(\sqrt{\bar{\alpha}(\tau)}X_0, (1 - \bar{\alpha}(\tau))\mathbf{I}\right), \quad \text{with} \quad \bar{\alpha}(\tau) := \exp\left(-\int_0^\tau \beta(u)du\right), \quad (28)$$

which matches Equation (23) at the grid points if we choose $\bar{\alpha}(\tau_k) = \bar{\alpha}_k$ (Song et al., 2021).

C.2 REVERSE-TIME DYNAMICS

We now reverse time to generate from noise to data. Let $\bar{\tau} = 1 - \tau$ denote the *generative time*. By classical time reversal diffusion (Anderson, 1982) the reverse-time process satisfies

$$dX_\tau = \left(-\frac{1}{2}\beta_\tau X_\tau - \beta(\tau)\nabla_X \log p_\tau(X_\tau)\right)d\bar{\tau} + \sqrt{\beta(\tau)}d\bar{B}_\tau, \quad \text{with} \quad d\bar{\tau} = -d\tau, \quad (29)$$

where p_τ are the forward-time marginals and $\nabla_X \log p_\tau$ is the score (Song et al., 2021).

In practice, most diffusion architectures parameterize the model via *noise prediction* ϵ_θ (Ho et al., 2020; Karras et al., 2022; 2024), which is related to the score by:

$$\nabla_X \log p_\tau(x) = -\frac{\epsilon_\theta(x, \tau)}{\sqrt{1 - \bar{\alpha}(\tau)}}. \quad (30)$$

C.3 TIME CHANGE TO FM

To embed VP diffusion into FM, we reparameterize time so that FM runs from noise to data, setting $t := 1 - \tau$, which yields the following FM schedules:

$$\alpha_t^{\text{FM}} := \sqrt{\bar{\alpha}(1-t)}, \quad \text{and} \quad \beta_t^{\text{FM}} := \sqrt{1 - \bar{\alpha}(1-t)}. \quad (31)$$

C.4 SCORE RELATIONS

For linear Gaussian reference paths, the score $s(x, t) := \nabla_X \log p_t(x)$ and the FM vector field $v_\theta(x, t)$ are linked by a schedule-dependent affine map (Lipman et al., 2023; Albergo et al., 2023; Liu et al., 2023):

$$s(x, t) = \frac{1}{\eta_t}\left(v_\theta(x, t) - \kappa_t x\right), \quad \kappa_t := \frac{\dot{\alpha}_t^{\text{FM}}}{\alpha_t^{\text{FM}}}, \quad \eta_t := \beta_t^{\text{FM}}\left(\frac{\dot{\alpha}_t^{\text{FM}}}{\alpha_t^{\text{FM}}}\beta_t^{\text{FM}} - \dot{\beta}_t^{\text{FM}}\right). \quad (32)$$

Combining the noise–score relation with the time change $\tau = 1 - t$ gives:

$$s(x, t) = \nabla_X \log p_t(x) = -\frac{\epsilon_\theta(x, 1-t)}{\beta_t^{\text{FM}}}, \quad (33)$$

since $\beta_t^{\text{FM}} = \sqrt{1 - \bar{\alpha}(1-t)}$. Substituting this into the score–velocity map yields the corresponding FM *velocity prediction* induced by an ϵ -parameterized diffusion model:

$$v_\theta(x, t) = \kappa_t x - \eta_t \frac{\epsilon_\theta(x, 1-t)}{\beta_t^{\text{FM}}}. \quad (34)$$

This identity lets an ϵ -trained diffusion model be used directly as an FM velocity field for the VP-induced schedules above; plugging v_θ into the FM SDE recovers the reverse-time VP sampler (and setting $\sigma \equiv 0$ recovers the probability-flow/DDIM ODE) under the change of variables $t = 1 - \tau$.

D HYPERPARAMETERS

D.1 SAMPLING PARAMETERS

For Stable Diffusion 3.5³ and FLUX.1 [dev]⁴, we follow the official sampling recommendations. Unless stated otherwise, we use the deterministic Euler scheduler with 28 inference steps for both models and generate images at 512×512 resolution. The classifier-free guidance scale is set to 4.5 for SD3.5 and 3.5 for FLUX. To ensure consistent extraction of cross-attention maps, we cap the maximum tokenized sequence length at 77 for SD3.5 and 256 for FLUX, and we verify that all prompts in our dataset fall within these limits. Models are loaded and all computations are performed in `bfloat16` to reduce memory usage.

D.2 METRICS

To summarize each hyperparameter setting with a single scalar, we macro-average the *relative improvement* over the base model across prompts, seeds, and metrics.

Let $X_{p,s}$ be the image produced by the current setting for prompt $p \in P$ and seed $s \in S$, and let $\hat{X}_{p,s}$ be the corresponding image from the base model. Let $\mathcal{M}$ denote the set of evaluation metrics. Since in our settings all metrics are increasing, the composite score for a hyperparameter setting is the macro-average

$$\frac{1}{|S|} \sum_{s \in S} \frac{1}{|P|} \sum_{p \in P} \frac{1}{|M|} \sum_{m \in M} \frac{m(X_{p,s}) - m(\hat{X}_{p,s})}{m(\hat{X}_{p,s})}, \quad (35)$$

such that a value larger than 0 indicates an average improvement over the base model, while values smaller than 0 indicate degradation.

D.3 TEST-TIME CONTROL

In the deterministic (ODE) variant, the single-pass update does not inherit the time-weighting $\frac{1}{2}\sigma_{\text{mem}}^2(t)$ that appears in the SDE case. Since $\sigma_{\text{mem}}(t)$ is large at early times and decays rapidly as $t \rightarrow 1$, we reintroduce this desirable early-strong / late-weak behavior in the ODE setting by reweighting the running cost:

$$f(X_t, t) = \lambda \cdot \sigma_{\text{mem}}^2(t) \cdot \text{Heuristic}(X_t), \quad (36)$$

where $\lambda > 0$ is the earlier introduced hyperparameter to account for different heuristic magnitudes. Throughout our test-time control experiments, we use this time-weighted running cost variant and sweep over $\lambda \in \{0.1, 0.5, 1, 2, 3, 4, 8, 12, 16, 32\}$. Values below 0.1 have negligible effect across heuristics, while values above 32 tend to produce artifacts (over-sharpening, texture noise) or occasional numerical instabilities (NaNs). See Figure 7 for qualitative trends.



Figure 7: Effect of the control parameter λ on test-time control with Stable Diffusion 3.5.

D.4 FINE-TUNING

We initialize the *memoryless* schedule from each model’s ODE 28-step inference schedule (same time steps), do not use classifier-free guidance, and for **FLUX.1** apply its native guidance scale (not CFG). Following Appendix D.1, we cap tokenized sequence length for cross-attention extraction to 77 (SD 3.5) and 256 (FLUX.1). Models are loaded in `bfloat16`; forward/backward passes run in

³<https://huggingface.co/stabilityai/stable-diffusion-3.5-medium>

⁴<https://huggingface.co/black-forest-labs/FLUX.1-dev>

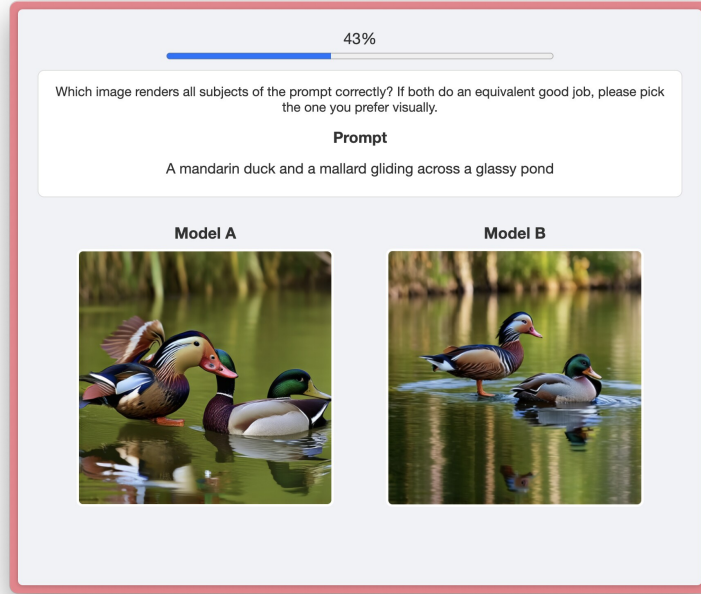


Figure 8: User interface for the prompt-conditioned, pairwise preference study.

BF16 and the final loss reduction is computed in FP32 to avoid numerical issues. To reduce memory, at each iteration we subsample 16 of the 28 steps to be used in our loss calculation. We further use a batch sizes of 5 trajectories for SD 3.5 and 2 trajectories for FLUX.1. We use two small prompt sets: HORSE&BEAR (single prompt: “A horse and a bear”) and TWOOBJECTS (15 prompts, each with two semantically similar subjects). Optimization uses AdamW with a weight decay of 0.01 and $\beta_0 = 0.95$, $\beta_1 = 0.999$. In addition, we also employ Accelerate to lower peak memory consumption. Table 5 lists the hyperparameter grids we sweep per heuristic; best settings are **bold**.

Table 5: Hyperparameter grids for fine-tuning; best settings per row in **bold**.

	Heuristic	Lambda λ	Learning rate	Checkpoint	Dataset
SD 3.5	Attend&Excite	{0.1, 1, 10 }	5e-5	{ 100 , 150 }	HORSE&BEAR
	CONFORM	{ 0.1 , 1, 10}	5e-5	{ 100 , 150 }	HORSE&BEAR
	Divide&Bind	{ 0.1 , 1, 10}	5e-5	{ 100 , 150 }	HORSE&BEAR
	FOCUS	{0.01, 0.1, 1 , 10, 100}	{1e-4, 5e-5 , 1e-5}	{ 100 , 150, 200}	{ HORSE&BEAR , TWOOBJECT}
FLUX.1	Attend&Excite	{0.1, 1, 10 }	5e-5	{ 200 , 250 }	HORSE&BEAR
	CONFORM	{0.1, 1, 10 }	5e-5	{ 200 , 250 }	HORSE&BEAR
	Divide&Bind	{ 0.1 , 1, 10}	5e-5	{ 200 , 250 }	HORSE&BEAR
	FOCUS	{0.01, 0.1, 1, 10, 100 }	{1e-4, 5e-5 , 1e-5}	{200, 250 , 300}	{ HORSE&BEAR , TWOOBJECT}

D.5 ADDITIONAL METRIC: OPEN-VOCABULARY DETECTION

As a complementary metric, we assess *subject presence* with OWL-V2 open-vocabulary detection (Minderer et al., 2023). For each prompt, we pass the subject strings as class queries and count an image as correct if *all* subjects are detected at least once. We report the fraction of images meeting this criterion.

Results for test-time control and fine-tuned models are shown in Tables 6 and 7. Both control algorithms increase subject presence over the base model. However, OWL-V2 is blind to attribute leakage and subject numerosity (it does not verify attributes or counts), so we exclude it from the main evaluation and report it only as a supportive metric here.

Table 6: OWL-V2 subject presence in percentage for test-time control.

Heuristic	SD3.5	FLUX
Base	69.33%	66.93%
Attend&Excite	72.13%	66.80%
CONFORM	77.20%	67.87%
Divide&Bind	70.80%	68.53%
FOCUS (Ours)	74.27%	68.27%

Table 7: OWL-V2 subject presence in percentage for fine-tuned models.

Heuristic	SD3.5	FLUX
Base	69.33%	66.93%
Attend&Excite	80.40%	74.93%
CONFORM	77.73%	72.53%
Divide&Bind	73.33%	63.87%
FOCUS (Ours)	78.53%	74.66%

E HUMAN STUDY

We evaluate whether metric gains translate to human preferences. Fifty participants each completed 40 prompt-conditioned, pairwise trials, resulting in 2,000 total judgments. In every trial, two images generated from the *same* prompt were shown side by side with the prompt; participants selected the image that better matched the prompt. The instruction shown was:

Which image renders all subjects of the prompt correctly? If both do an equivalent good job, please pick the one you prefer visually.

To ensure sufficient rating density, we fixed the sampling seed to 0, yielding one image per method–prompt pair (pool of 150 prompts). Trials were balanced across backbone and setting: SD 3.5 vs. FLUX.1 and test-time control vs. fine-tuning each accounted for one quarter of the comparisons per participant. A screenshot of the interface is shown in Figure 8.

E.1 ELO RATING COMPUTATION

We compute Elo ratings from the pairwise outcomes to obtain an across-method ranking, alongside win rates (fraction of pairwise wins). Elo is initialized at 1500 for all candidates and updated after each comparison with $K=32$. For a candidate A with rating R_A matched against B with R_B , the expected score is

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}, \quad (37)$$

and the update is

$$R'_A = R_A + K(S_A - E_A) \quad (38)$$

where $S_A = 1$ for a win, 0 for a loss, and 0.5 for a draw. Higher Elo indicates stronger preference relative to alternatives. Win rate is reported as the proportion of head-to-head wins.

F EXTRA SAMPLES

F.1 TEST-TIME CONTROL: STABLE DIFFUSION 3.5

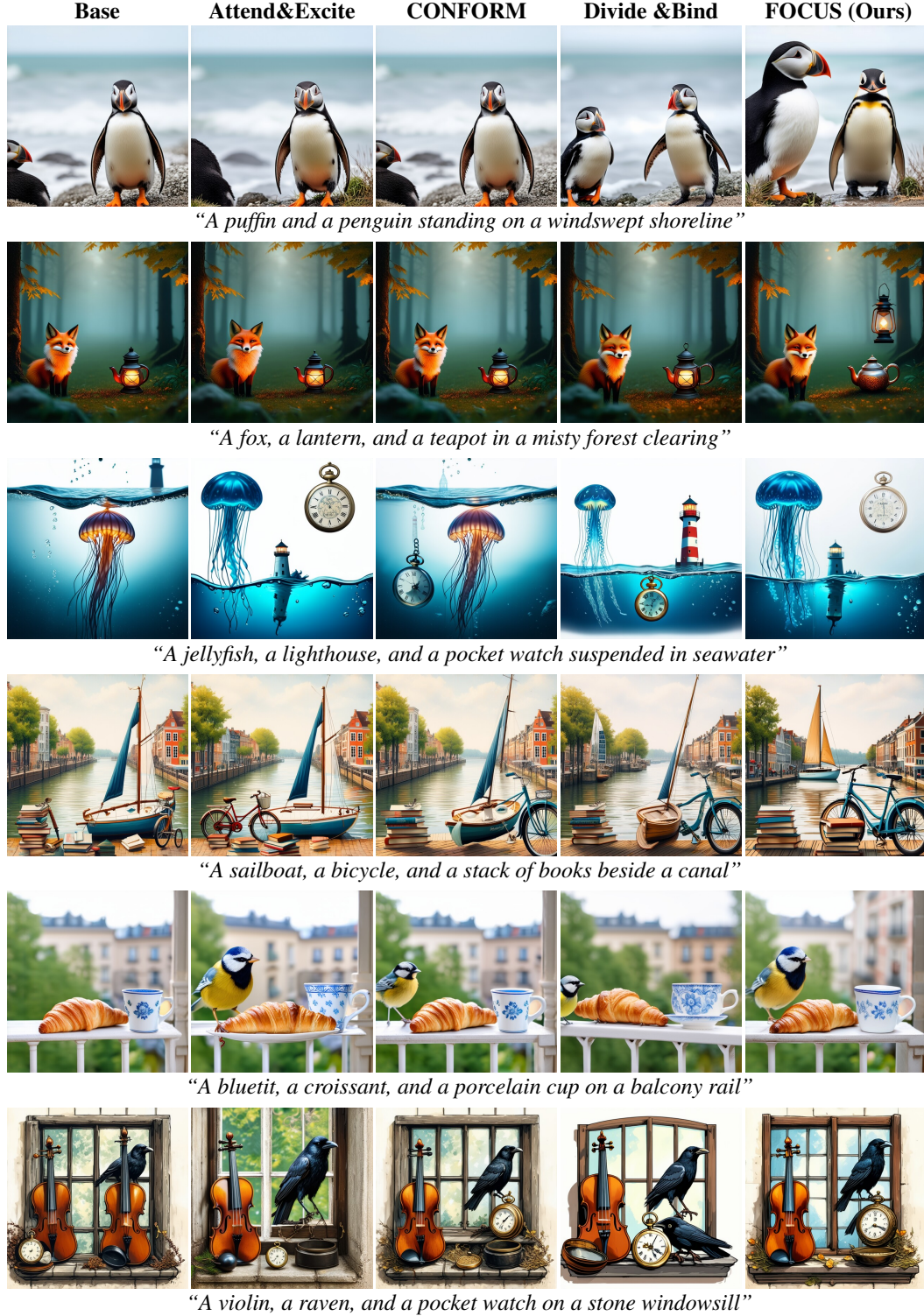


Figure 9: Stable Diffusion 3.5 samples with test-time control. All evaluated heuristics are shown at their optimal λ .

F.2 TEST-TIME CONTROL: FLUX.1 [DEV]



Figure 10: FLUX.1 dev samples with test-time control. All evaluated heuristics are shown at their optimal λ .

F.3 FINE-TUNED: STABLE DIFFUSION 3.5

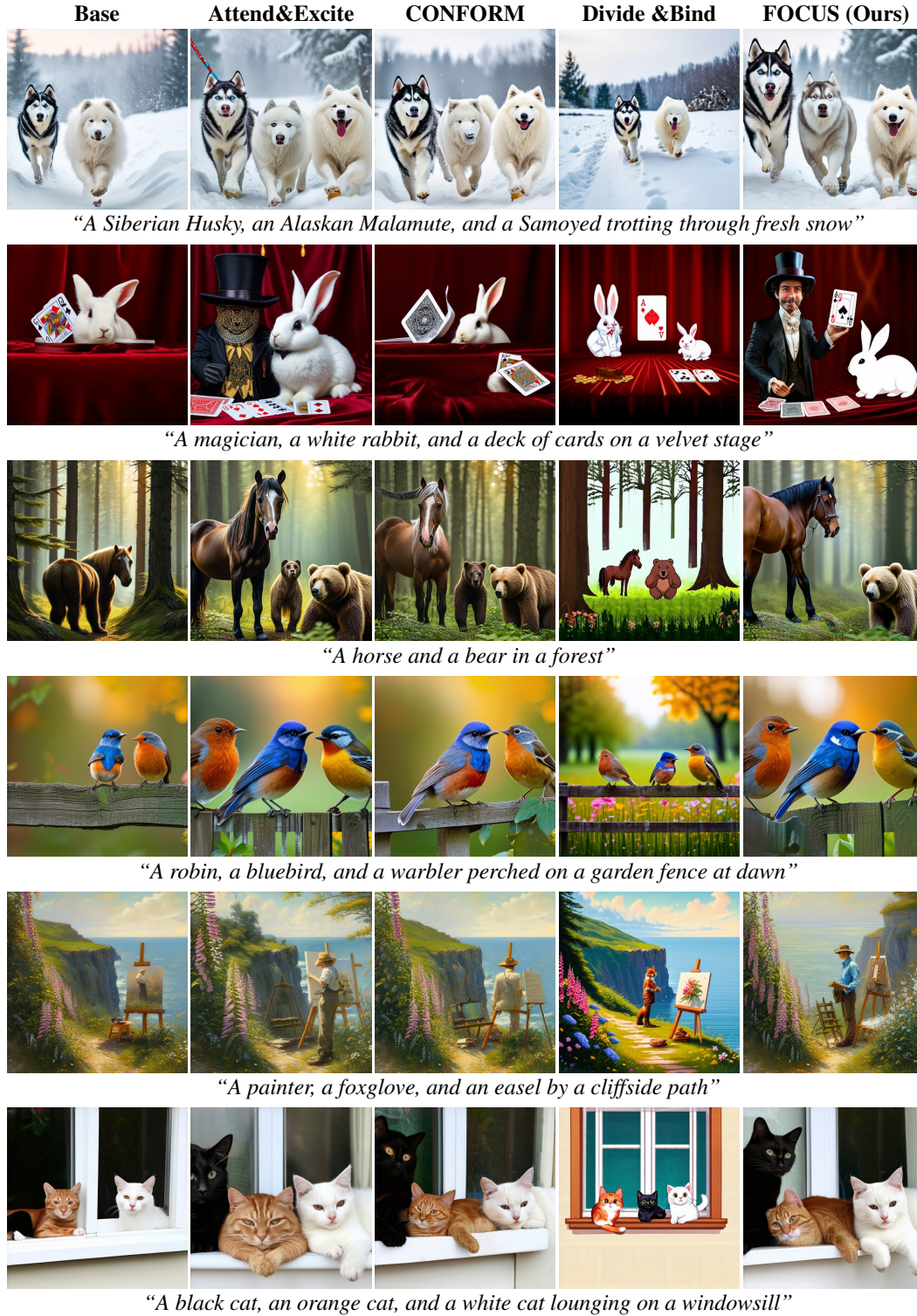


Figure 11: Sample results from Stable Diffusion 3.5 fine-tuned with each heuristic. Prompts were not seen during training to evaluate generalization. All images are generated with identical settings (seed, sampler, steps, guidance); each heuristic is shown at its optimal trained λ .

F.4 FINE-TUNED: FLUX.1 [DEV]

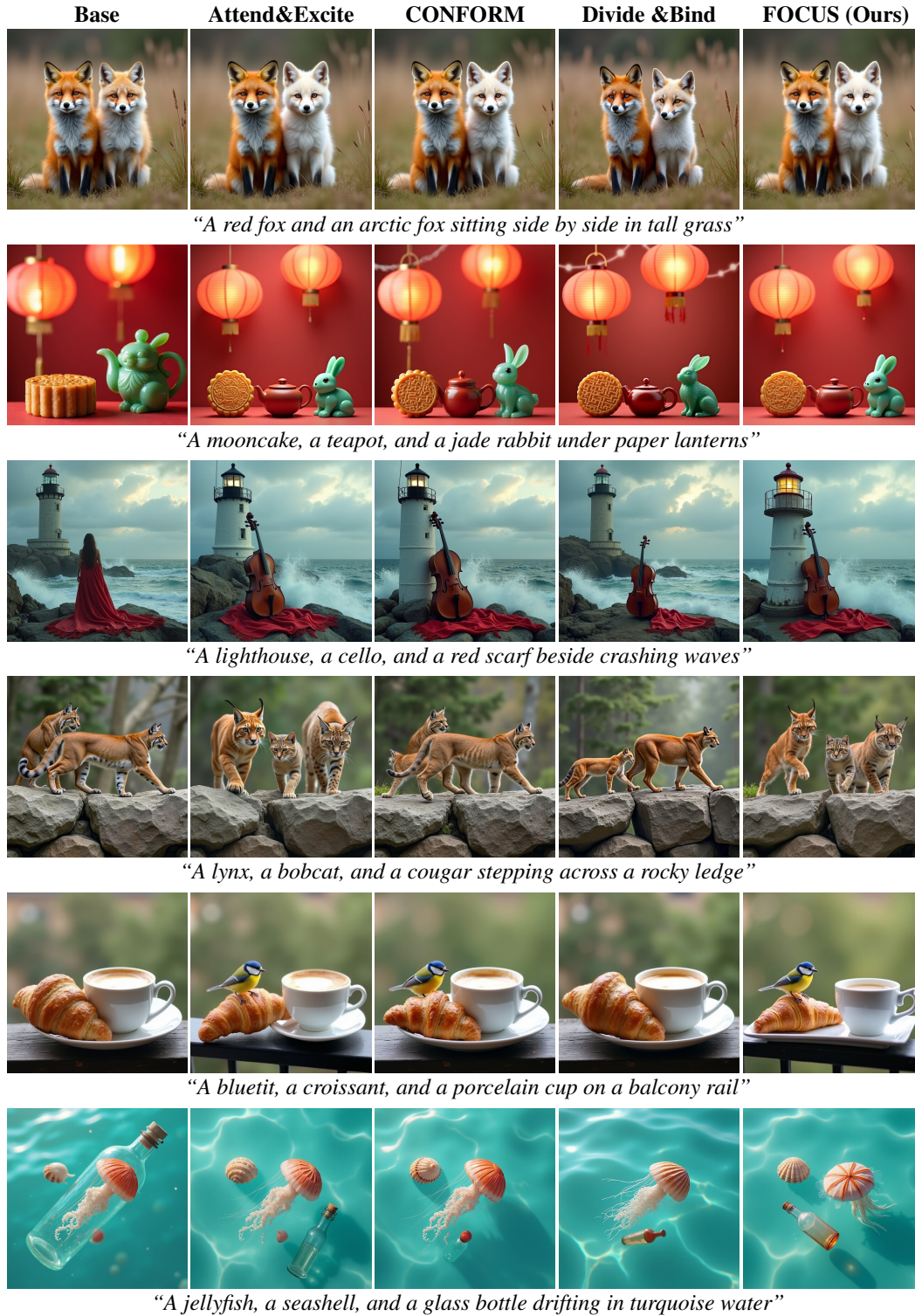


Figure 12: Sample results from FLUX.1 [dev] fine-tuned with each heuristic. Prompts were not seen during training to evaluate generalization. All images are generated with identical settings (seed, sampler, steps, guidance); each heuristic is shown at its optimal trained λ .