POSITION-SENSITIVE CONCEPT ACQUISITION IN TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained text-to-image models possess rich visual priors, yet reliable control over quantitative attributes can remain difficult to acquire through supervised fine-tuning. We study how the position of a target condition affects this acquisition process. Across models, concepts, and evaluation settings, paired fine-tuning with identical image supervision and matched training budgets yields markedly different control capabilities when the same target field is placed at the beginning or end of the prompt. The effect depends on both the model and the concept. Using angle control with Qwen-Image as a representative case, we find evidence linking the acquisition gap to information propagation in the frozen causal text encoder and the use of target information during training. A simple Prefix fine-tuning scheme repeats the target condition at the beginning and improves control over standard end-position training when the prefix is retained at inference. We further introduce a pre-adaptation probe of relative preference for correct over counterfactual conditions. Cross-model comparisons and controlled pretraining experiments illustrate three characteristic acquisition patterns under finite budgets: weak control at both positions, a pronounced positional gap, and stronger control at both positions. These patterns are associated with initial support and learning progress, rather than fixed support thresholds. Acquired control may remain format-dependent even as positional gaps narrow.

1 INTRODUCTION

Pretrained text-to-image models generate rich visual content (Rombach et al., 2022; Saharia et al., 2022), yet fine-grained prompt adherence remains challenging (Ghosh et al., 2023; Huang et al., 2023). We study precise control of angles, lengths, and proportions through supervised fine-tuning, aiming to establish reliable text–attribute correspondences without assuming these concepts are entirely novel to the model. Even explicit and consistent image–text supervision does not guarantee such control within a finite training budget, motivating a closer examination of what governs its acquisition.

We identify a pronounced phenomenon of **position-sensitive concept acquisition** in text-to-image models. In the angle control task with Qwen-Image (Wu et al., 2025), changing the placement of the angle field while preserving target semantics produces markedly different outcomes under fixed training images, target angles, and optimization budgets. Figure 1(a) illustrates the paired comparison: separately fine-tuned models are tested in their respective training formats, with Front achieving stronger angle control than Back. Beyond the semantics of a condition, its placement in the prompt can therefore affect the control established during fine-tuning.

Prior work on condition position has primarily examined positional effects at inference, including how token positions influence text representations, attribute binding, and the resulting images (Ma et al., 2024; Abbasi et al., 2025). Other studies have investigated how training captions or prompt structure affect adaptation (Merchant et al., 2025). In contrast, we focus on concept acquisition during training: when image supervision, target semantics, and optimization budgets are held constant, does the position of the same target condition affect the acquisition of control? Our question thus concerns not only how existing capabilities are invoked at inference, but also whether new control capabilities can be established through training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To examine the generality of this phenomenon, we extend our experiments to additional quantitative concepts, including differences between ellipse axis lengths and progress bar proportions, and to other text-to-image models, including Z-Image (Z-Image Team et al., 2025) and FLUX.2 Klein (Black Forest Labs, 2026). These experiments show that positional effects depend strongly on both the model and the concept. Figure 1(b) summarizes possible acquisition patterns across settings. Motivated by this observation, we investigate two questions: **How does condition position affect the use of target information during fine-tuning, and why does position sensitivity vary across models and concepts?**

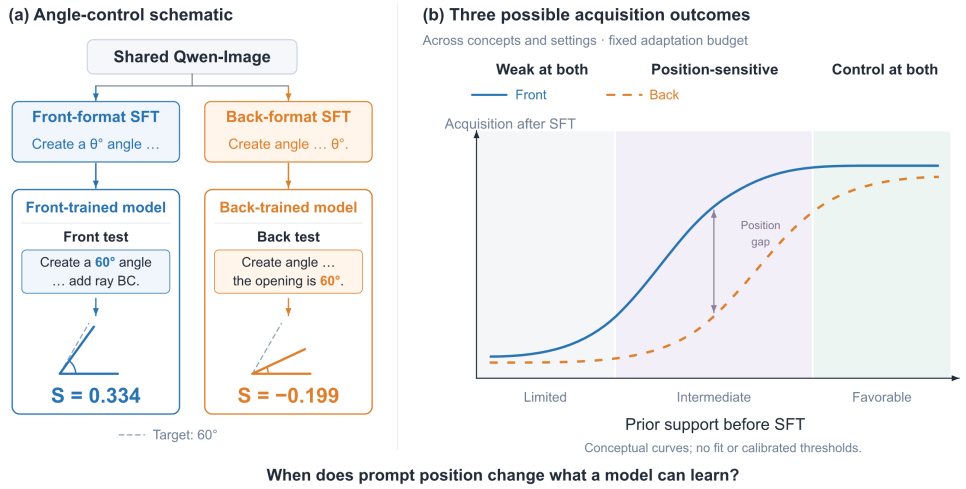

Figure 1: **Position-sensitive acquisition and a qualitative interpretation.** (a) Two models start from the same Qwen-Image checkpoint and undergo Front- or Back-format SFT with matched supervision and budgets. Each is tested in its own training format. Displayed scores are aggregate angle-control results on Different/Out, $A_{FF} = 0.334$ and $A_{BB} = -0.199$, rounded from Table 1; they are not measurements of the illustrated 60° example. Angle sketches and abbreviated prompts are schematic, not generated samples or verbatim experimental templates. The score S measures normalized MAE control, not accuracy (Eq. 1). (b) Three possible outcomes across concepts and settings under a fixed adaptation budget. The conceptual curves are not fitted and imply no universal monotonic prior–control relationship, calibrated thresholds, or model-specific prior locations.

First, we analyze the positional effect using the Qwen-Image angle task as a representative case. Placing the angle condition at the beginning influences a broader set of contextual representations in the frozen causal text encoder and induces stronger condition-related responses in the generator. Building on this analysis, we introduce **Prefix fine-tuning**: we repeat the target angle at the beginning while retaining its original description at the end, and train using this format. With the prefix retained at inference, this substantially improves angle control over standard end-position fine-tuning. Removing it reduces performance, consistent with a change in the use of conditioning information rather than elimination of format dependence.

Second, to investigate variation across models and concepts, we propose a **pre-adaptation probe of prior support for concept acquisition**. Using the frozen pretrained model's native prediction objective, the probe compares prediction errors under correct and counterfactual conditions. Relative preference for the correct condition characterizes initial discriminative support for the target concept. Comparisons suggest an association between this support and subsequent control, but similar probe scores can accompany different acquisition patterns; the probe is not a general predictor of fine-tuning outcomes.

To further examine initial support and position sensitivity, we construct a controlled shape generation environment and vary target-shape exposure during visual pretraining. With architecture and training setup fixed, paired fine-tuning illustrates **three empirical acquisition patterns**: weak control at both positions, a pronounced positional gap, and stronger control at both positions. These patterns depend on measured initial support and learning progress under a finite budget, rather than fixed exposure

thresholds. A single exposure setting can show an early gap that narrows with further training, without implying that every model traverses the same sequence.

Our main contributions are as follows:

1. **We identify position-sensitive concept acquisition during training** through paired fine-tuning experiments across models and concepts, showing that condition placement affects quantitative control under matched supervision and optimization budgets.
2. **We connect condition position to target information propagation and use** through a representative Qwen-Image angle study, and introduce Prefix fine-tuning to improve control for prompts with end-position conditions.
3. **We introduce a pre-adaptation probe of concept support and characterize three empirical acquisition patterns** through cross-model comparisons and controlled pretraining experiments, relating position sensitivity to initial support and learning progress.

2 Related Work

2.1 Text Conditioning Representations and Positional Effects

Text conditioning draws on representations learned through image–text contrastive training, as in CLIP (Radford et al., 2021), or text-only pretraining, as demonstrated by Imagen (Saharia et al., 2022).

Diffusion Lens examines concepts across text encoder depths (Toker et al., 2024), while Magnet diagnoses embedding bias and entanglement and improves attribute binding (Zhuang et al., 2024). DAAM analyzes text–region associations through cross-attention (Tang et al., 2023), and Prompt-to-Prompt controls these maps for text-guided editing (Hertz et al., 2023). Structured Diffusion Guidance, Attend-and-Excite, and CONFORM address concept omission and attribute binding in multi-object generation (Feng et al., 2023; Chefer et al., 2023; Meral et al., 2024).

Ma et al. (2024) analyze positional biases in decoder-only prompt encoders. Abbasi et al. (2025) trace CLIP's early-object preference to training and examine its effect on generated object salience. Consistent caption structure improves alignment relative to shuffled fields during fine-tuning (Merchant et al., 2025). Our paired Front/Back experiments preserve within-format structure, supervision, and budgets to examine how target-field position affects quantitative control acquisition.

2.2 Concept Learning Dynamics and Probing Prior Support

DreamBooth fine-tunes a generator to bind a subject to an identifier (Ruiz et al., 2023), whereas Textual Inversion learns concept embeddings with the model frozen (Gal et al., 2023). Our focus is quantitative control across target values, rather than subject or style personalization.

Controlled synthetic tasks link concept frequency to compositional learning (Okawa et al., 2023). Park et al. (2024) relate concept signals to learning speed and order and identify latent capabilities not elicited by standard prompts. Yang et al. (2025b) explain non-monotonic test-loss dynamics in compositional generalization. Concept-TRAK attributes generated concepts to training examples using concept-specific losses (Park et al., 2026).

Pretrained diffusion models also support zero-shot classification through conditional likelihood or denoising comparisons (Li et al., 2023; Clark & Jaini, 2023). Our pre-adaptation probe uses native prediction errors to compare correct and counterfactual conditions, then relates initial discrimination to subsequent control and position sensitivity under finite training budgets. It is not a classifier's top-1 accuracy or a guarantee of successful adaptation.

3 Systematic Validation of Position-Sensitive Concept Acquisition

We evaluate position-sensitive concept acquisition across models, concepts, prompt formats, and training durations under matched supervision and budgets.

3.1 Experimental Setup and Evaluation

We operationally define concept acquisition as establishing a reliable correspondence between textual target values and generated visual attributes. This concerns acquiring controllable mappings, not establishing that the underlying concept was absent from the pretrained model. The tasks specify an angle, the difference between an ellipse's vertical and horizontal axis lengths, or a progress bar's filled proportion.

Paired training. For a target value p and template τ, prompts $c_F(p, \tau)$ and $c_B(p, \tau)$ place the target condition at the beginning (Front, F) and end (Back, B), respectively, with grammatical adjustments where needed. The comparison preserves target semantics and supervision but is not a strictly token-identical reordering. Paired runs share the same training images, target values, and pretrained checkpoint, with matched dataset sizes, optimization settings, and budgets. The text encoder remains frozen, while the generator is updated using each model's adaptation configuration; position comparisons are always made within the same model-concept pair. For the geometric tasks with Qwen-Image, each position uses 1,000 image-text pairs, a batch size of 4, and 8 epochs by default.

Control metric. For target values p_i and attributes $\hat{p}_i$ measured from generated images, we normalize mean absolute error by that of the best constant predictor on the same test set:

$$S = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} |\hat{p}_i - p_i|}{\min_a \frac{1}{N}\sum_{i=1}^{N} |a - p_i|}. \tag{1}$$

Here, $S = 1$ indicates exact control and $S = 0$ the best constant baseline; negative scores indicate greater error, and positive scores alone do not establish full acquisition. For example, $S = 0.3339$ means a 33.39% reduction in MAE relative to the constant baseline, not classification accuracy. We denote this same score by A_{st} when specifying training position s and test position t.

Evaluation splits. For geometric tasks, Same/Different denotes seen/unseen target values, and In/Out denotes seen/unseen prompt templates. Unseen values lie within the training range, so this evaluation tests interpolation. Task definitions, template allocation, and the split table are given in Appendix A.1.

3.2 Position Sensitivity across Models and Concepts

Table 1 compares angle, ellipse axis-length difference, and progress bar control with Qwen-Image, Z-Image, and FLUX.2 Klein 4B. Geometric tasks use Different/Out, with unseen target values and templates; progress bar control uses a separately constructed test set.

Table 1: Control scores across models and concepts. The first and second subscripts denote training and test positions, respectively. Ellipse denotes the axis-length difference and progress bar denotes the filled proportion. Geometric tasks use Different/Out; progress bar control uses a separate test set.

Model	Concept	A_{FF}	A_{FB}	A_{BF}	A_{BB}
Qwen-Image	Angle	0.3339	−0.0975	−0.1454	−0.1986
Qwen-Image	Ellipse	0.5810	−0.2010	−0.5278	−0.2853
Qwen-Image	Progress bar	0.6164	0.4518	0.4700	0.5027
Z-Image	Angle	0.0241	−0.0686	−0.0938	−0.0821
Z-Image	Ellipse	0.5624	−0.2373	−0.0348	0.2915
Z-Image	Progress bar	−0.4400	−0.5236	−0.5391	−0.4745
FLUX.2 Klein 4B	Angle	−0.2812	−0.3608	−0.3815	−0.4070
FLUX.2 Klein 4B	Ellipse	0.0764	−0.9868	−0.3804	−0.1469
FLUX.2 Klein 4B	Progress bar	−0.6236	−0.6109	−0.6891	−0.5164

When training and test positions match, three patterns emerge. Qwen-Image exhibits pronounced gaps on angle and ellipse control, with positive front-position and negative back-position scores;

Figure 1(a) illustrates the angle comparison. Qwen-Image on progress bar control and Z-Image on ellipse control outperform the constant baseline at both positions. Z-Image on angle and progress bar control, together with FLUX.2 Klein 4B on the tested concepts, exhibits weak overall control at both positions. The advantage of front placement therefore depends on both the model and the concept.

3.3 Training Dependence and Robustness

Comparing A_{FF} with A_{BB} changes both training and test positions. To assess the effect of training position under a fixed test format, we compare A_{FF} with A_{BF} for front-position tests and A_{FB} with A_{BB} for back-position tests.

For angle control with Qwen-Image, fixing the test condition at the beginning yields $A_{FF} = 0.3339$ after front-position training and $A_{BF} = -0.1454$ after back-position training. Moving the condition to the beginning only at inference therefore does not recover the control achieved through front-position training. With the test condition fixed at the end, Front training also scores higher ($A_{FB} = -0.0975$ versus $A_{BB} = -0.1986$), but both remain below the constant baseline.

A separate comparison holds the trained model fixed: switching the front-trained model from Front to Back testing reduces its score from $A_{FF} = 0.3339$ to $A_{FB} = -0.0975$. Training position thus affects attainable control, but the improvement does not automatically transfer across formats. This dependence varies across tasks: Z-Image ellipse scores are positive under FF and BB but negative under FB and BF, whereas Qwen-Image progress bar scores remain positive under all four combinations.

Robustness checks. The Qwen-Image angle gap persists across all four target-value/template splits, while back-position ellipse control yields positive scores on seen values but negative scores on unseen values within the training range (Appendix A.2). The angle gap also persists at 4, 8, and 16 epochs (Appendix A.3). Its narrowing from 8 to 16 epochs reflects both improved back-position and reduced front-position scores, and cannot be attributed solely to the former catching up. These results establish persistence within the tested budgets, without implying that back-position training can never acquire control.

4 Mechanistic Analysis of Condition Position and an Intervention

Using the Qwen-Image angle task illustrated in Figure 1(a), we examine how condition position affects information propagation and its use during training, then evaluate a Prefix intervention.

4.1 Condition Position Affects the Propagation of Target Information

With causal self-attention (Vaswani et al., 2017), an early condition can affect subsequent text representations, whereas a late condition cannot affect preceding tokens. The generator can still access the late angle token directly. We therefore ask whether front placement distributes target information more broadly across the conditioning representations and elicits a stronger generator response.

For front-position and back-position prompts, we replace the correct angle p with a counterfactual $p' \neq p$, keeping the remaining content unchanged within each format. We measure changes in the angle token, context spread (average change across other tokens), coverage (the fraction exceeding the probe threshold), and generator response (relative L_2 prediction change). Generator comparisons hold the image latent and noise variables fixed; further probe details are given in Appendix B.1.

Before fine-tuning, angle tokens respond at both positions, but front placement produces approximately 3.74 times the context spread and a stronger generator response (Table 2). Broader propagation persists across all four target-value/template splits. These observations are consistent with the causal-encoding asymmetry, while leaving open how the generator learns to use the available information.

Table 2: Propagation and prediction responses before fine-tuning on the Qwen-Image angle task. Context spread and coverage are measured at text encoder layer 28.

Metric	Front condition	Back condition
Target token representation change	0.231842	0.211280
Context spread	0.038934	0.010405
Context coverage	0.503627	0.156021
Relative L_2 change in generator predictions	0.016406	0.012576

4.2 Use of Conditioning Information during Training

To distinguish image fitting from condition discrimination during fine-tuning, let $\ell_k^R(x, c; \omega)$ denote checkpoint k's native prediction loss within the angle region of interest (ROI), with noise and time variables ω shared across paired conditions. The counterfactual loss gap at position s is

$$M_k^R(s) = \mathbb{E}\left[\ell_k^R\big(x, c_s(p', \tau); \omega\big) - \ell_k^R\big(x, c_s(p, \tau); \omega\big)\right]. \tag{2}$$

For each position s, checkpoint k belongs to that position's own training trajectory and is probed in the matching format; comparisons therefore involve separately trained models. The correct-condition loss measures image fitting, while $M_k^R(s) > 0$ indicates a preference for the correct angle. The relative L_2 prediction response measures sensitivity to changing the condition, but not whether that change is correct. Generation control in Section 3 tests the further requirement that generated attributes follow the requested values.

Both training positions improve image fitting and condition discrimination, yet back-position training still yields weak generation control. Lower loss and stronger discrimination therefore do not directly guarantee reliable numerical control.

At epoch 8, the prediction response is stronger for the front condition, with a paired 95% confidence interval for the difference that excludes zero. In contrast, the interval for the counterfactual loss-gap difference includes zero. These intervals quantify sampling uncertainty in the paired probe evaluation, not variability across independent SFT runs. Numerical diagnostics, confidence intervals, and background-region results are reported in Appendix B.2.

4.3 Prefix Fine-Tuning and Its Effects

Motivated by broader propagation, Prefix fine-tuning uses $c_P(p, \tau) = \pi(p) \oplus c_B(p, \tau)$, where $\pi(p)$ repeats the target angle at the beginning and $\oplus$ denotes text concatenation. Unlike continuous prefix-tuning (Li & Liang, 2021), we repeat discrete text and update the generator, without learning prefix vectors. For example, a prompt ending in "the angle is 20°" receives "20°." at the beginning. Training uses the same 1,000 images and target values. Table 3 compares standard, mixed, and Prefix training across the four splits; the Prefix rows evaluate the same models with the prefix retained or removed. Mixed-training configuration and supplementary probes are given in Appendix B.3.

Table 3: Angle control under different training schemes. Prefix rows use the same models, averaged over SFT seeds 42 and 44. "Retained" keeps both the prefix and the end-position description; "removed" keeps only the latter.

Training	Test format	Same/In	Same/Out	Different/In	Different/Out
Front	Front	0.4750	0.3912	0.3297	0.3339
Back	Back	−0.1293	−0.2448	−0.1162	−0.1986
Mixed	Front	−0.0620	−0.0395	−0.1286	−0.0331
Mixed	Back	−0.0033	−0.1723	−0.1675	−0.1499
Prefix	Retained	0.5211	0.5396	0.2184	0.2369
Prefix	Removed	−0.1211	−0.0368	−0.0908	−0.0486

With the prefix retained, control scores are positive and exceed standard Back on all four splits. Prefix outperforms Front on seen target values but remains weaker on unseen values. Removing the prefix makes all four scores negative; the current Mixed configuration also remains below the constant baseline at both test positions.

In a separate 8-epoch comparison, Back and Prefix have similar correct-condition ROI losses, but the counterfactual loss gap increases from 0.000306 to 0.001211 with Prefix. Thus, improved generation control accompanies stronger condition discrimination. Selective perturbations further show greater sensitivity to prefix replacement after Prefix training and to suffix replacement after Back training (Appendix B.3), consistent with dependence on the learned conditioning format.

The removal results indicate that the Prefix training format should be retained at inference. Prefix also repeats the target value and increases prompt length; the current interventions do not fully separate these factors from propagation extent and do not isolate a single causal mechanism.

Propagation provides a mechanistic clue for this angle task, but alone does not explain why Qwen-Image learns progress bar control at both positions. We next examine whether support already present before adaptation helps account for such model–concept differences.

5 Prior Support, Learning Progress, and Position Sensitivity

We next ask whether prior support helps explain variation in position sensitivity across models and concepts, using a pre-adaptation probe and controlled experiments that vary concept exposure and track learning progress.

5.1 Probing Prior Support for Concept Acquisition

Section 4.2 tracks counterfactual loss gaps during adaptation. Here, we instead measure relative preferences before any updates, using the frozen pretrained model. For each target image, we construct $K = 8$ prompts: one with the correct target value and seven replacing only that value. Candidates share the image, noise realization, timestep, template, and remaining descriptions. Using each model's native prediction objective, let L_* and L_j denote errors under the correct and counterfactual conditions. The prior score is the pairwise discrimination accuracy:

$$P = \mathbb{E}\left[\frac{1}{K-1}\sum_{j\neq *}\left(\mathbf{1}[L_* < L_j] + \frac{1}{2}\mathbf{1}[L_* = L_j]\right)\right], \qquad K = 8. \tag{3}$$

P measures how often the correct condition has lower error. It is a pairwise win rate, not eight-way classification accuracy, so its chance level is 0.5. Relative preferences avoid comparing raw loss scales, but do not by themselves calibrate difficulty across models or concepts. We report the All configuration, with background correction for Qwen-Image. This pre-adaptation measure does not use subsequent training outcomes and captures initial discriminative support, not generation control itself.

5.2 Prior Strength and Concept Acquisition

Table 4 relates prior scores to matched-position control after fine-tuning (A_{FF}: front; A_{BB}: back). Geometric tasks use Different/Out, with unseen target values and templates; progress bar control uses its dedicated test set. Prior scores are empirical pairwise discrimination accuracies computed on the corresponding pre-adaptation evaluation sets.

Within Qwen-Image, progress bar control combines stronger initial discrimination with above-baseline control at both positions, whereas ellipse control has weaker initial support and a pronounced positional gap. On progress bars, Qwen-Image also has a higher prior score and better control than Z-Image under their respective adaptation settings.

The association is not a predictive rule: Qwen-Image angle ($P = 0.5125$) and Z-Image ellipse ($P = 0.5179$) have similar prior scores, yet the former has a pronounced matched-position gap and the latter positive scores at both positions. Similar probe readings can therefore accompany different

Table 4: Pre-adaptation prior scores and matched-position control after fine-tuning. Prior scores report measured pairwise discrimination accuracy. Ellipse and progress bar abbreviate the axis-length difference and filled-proportion tasks.

Model	Concept	Prior score P	A_{FF}	A_{BB}
Qwen-Image	Angle	0.5125	0.3339	−0.1986
Qwen-Image	Ellipse	0.5518	0.5810	−0.2853
Qwen-Image	Progress bar	0.6250	0.6164	0.5027
Z-Image	Angle	0.5429	0.0241	−0.0821
Z-Image	Ellipse	0.5179	0.5624	0.2915
Z-Image	Progress bar	0.5446	−0.4400	−0.4745

acquisition patterns. We next vary visual exposure while fixing the architecture and training setup to reduce the confounds of cross-model comparisons, rather than to establish a universal threshold for P.

5.3 Controlled Pretraining Reveals an Acquisition Regime of Position Sensitivity

To isolate prior support from model- and task-specific confounds, we construct a controlled shape-generation environment in which target-concept exposure can be manipulated directly. The model uses a frozen Qwen3-0.6B-Base causal text encoder (Yang et al., 2025a) and a ∼28M-parameter conditional U-Net trained in pixel space. Prompts describe shape, color, and size. Six shapes serve as base concepts, while star, heart, and trapezoid are designated as target concepts. During 10,000-step pretraining, each target shape is sampled with probability $\alpha/9$, while architecture, training budget, and the remaining data distribution are fixed. We sweep 16 exposure levels over $\alpha \in [0, 0.004]$. Importantly, α controls visual exposure of the generator and is therefore an intervention rather than a calibrated measure of prior strength; the pretrained text encoder remains unchanged.

From each pretrained checkpoint, we perform paired Front- and Back-format SFT for 100 steps on the same 18 target shape–color–size combinations. Paired runs share the initial checkpoint, image sequence, diffusion noise, timesteps, and optimization settings, differing only in whether the shape field appears first or last. We repeat SFT with three random seeds and evaluate 16 checkpoints from step 0 to 100. Further training and evaluation details are given in Appendix C.1.

Let $G_{FF}(\alpha, k)$ and $G_{BB}(\alpha, k)$ denote matched-position shape accuracies after Front and Back training, respectively, at exposure α and SFT step k. We define the matched-position acquisition gap as

$$\Delta(\alpha, k) = G_{FF}(\alpha, k) - G_{BB}(\alpha, k). \tag{4}$$

Figure 2 reports the mean Δ across three SFT seeds, conditional on pretraining seed 0. Positive values indicate stronger matched-position control after Front training. The gray and purple contours mark regions where mean G_{FF} is below 15% and above 90%, respectively; these thresholds are visualization aids rather than calibrated phase boundaries.

The resulting map shows that position sensitivity is concentrated in an intermediate acquisition regime. When learning is still weak, both formats perform poorly and the gap is generally small or unstable. Once the target concepts become learnable but before control saturates, a broad positive Front advantage emerges. At $\alpha = 0.002$, for example, the mean gap reaches 16.2 percentage points at step 10 and 19.4 points at step 20. At $\alpha = 0.003$, it reaches 23.6 points at step 25 and remains above 20 points through steps 35–40. With further training, the gap contracts as both formats approach high matched-position accuracy.

This behavior is also visible along a single training trajectory. At $\alpha = 0.003$, the matched-position gap changes from −6.9 percentage points before SFT to 23.6 at step 25, then falls to 1.9 by step 100. Position sensitivity therefore need not be inherited from an initial inference-position difference; it can emerge during acquisition and later diminish as learning progresses.

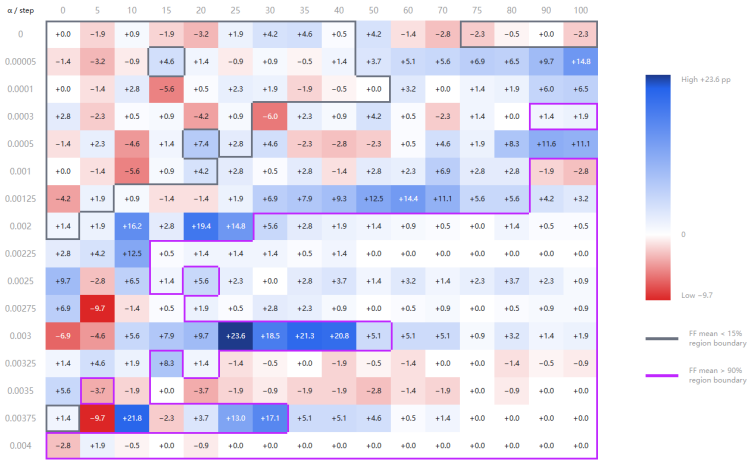

Figure 2: **Position sensitivity across controlled pretraining exposure and SFT progress.** Each cell shows the mean matched-position gap $\Delta = G_{FF} - G_{BB}$ in percentage points across three SFT seeds, conditional on pretraining seed 0. Rows vary target-shape exposure α, and columns track SFT progress. Gray and purple contours mark $G_{FF} < 15\%$ and $G_{FF} > 90\%$, respectively. The strongest Front advantages concentrate between the low-learning and saturated regions.

These results suggest that position sensitivity is neither a fixed property of a concept nor determined by a single exposure threshold. With insufficient support, neither position learns reliably; when control is already easy to acquire, both formats approach saturation and the observable gap shrinks. The largest differences instead occur while the controllable mapping is actively being acquired. The non-monotonic dependence on α further cautions against treating exposure itself as a calibrated prior-strength axis. Taken together, the controlled experiments support a dynamics-based view: prior support influences where learning begins, while condition position affects how rapidly usable control is acquired from that starting point.

6 Conclusion

Condition placement can substantially affect quantitative control acquired through fine-tuning under identical image supervision and matched budgets, with effects varying across model–concept pairs. Our Qwen-Image angle study provides evidence linking this variation to the propagation and use of conditioning information, motivating Prefix fine-tuning as a format-dependent intervention. Cross-model probes and controlled visual pretraining further relate acquisition to initial support and learning progress. Under finite budgets, we observe three characteristic patterns: weak control at both positions, a pronounced gap, and stronger control at both positions. These are empirical outcomes, not categories determined by a universal prior-score threshold or stages that all models necessarily traverse. A gap can narrow as learning proceeds while control remains dependent on the input format.

For tasks exhibiting position sensitivity, the results motivate evaluating front placement or Prefix fine-tuning and preserving the learned format at inference. Prefix improves on the Back baseline in our angle study but is not uniformly better than Front training, and its benefits do not isolate propagation from repetition or prompt length. Likewise, the pre-adaptation probe is informative about discrimination, not a guaranteed predictor of generation control. Practical assessment should therefore consider control quality, training-position differences, and cross-format transfer separately, within the model, task, and budget being adapted.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. CLIP under the microscope: A fine-grained analysis of multi-object representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2502.19842.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026.

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*, 42(4):148, 2023. doi: 10.1145/3592116. URL https://arxiv.org/abs/2301.13826.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b87bdcf963cad3d0b265fcb78ae7d11e-Abstract-Conference.html.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.05032.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract.html.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.16203.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.11831.

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*, pp. 9005–9014, 2024. URL https://conform-diffusion.github.io/.

Nicholas Merchant, Haitz Sáez de Ocáriz Borde, Andrei Cristian Popescu, and Carlos Garcia Jurado Suarez. Structured captions improve prompt adherence in text-to-image models (Re-LAION-Caption 19M). *arXiv preprint arXiv:2507.05300*, 2025. URL https://arxiv.org/abs/2507.05300.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d0f188c7947eacb0c07f709576824f6-Abstract-Conference.html.

Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.19370.

Yonghyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Naoki Murata, Wei-Hsiang Liao, Woosung Choi, Kin Wai Cheuk, Junghyun Koo, and Yuki Mitsufuji. Concept-TRAK: Understanding how diffusion models learn concepts through concept-level attribution. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.06547.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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*, 2022. URL https://arxiv.org/abs/2112.10752.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. URL https://arxiv.org/abs/2208.12242.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract.html.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5644–5659, 2023. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9713–9728, 2024. doi: 10.18653/v1/2024.acl-long.524. URL https://aclanthology.org/2024.acl-long.524/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming 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. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,

Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Yongyi Yang, Core Francisco Park, Ekdeep Singh Lubana, Maya Okawa, Wei Hu, and Hidenori Tanaka. Swing-by dynamics in concept learning and compositional generalization. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2410.08309.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2409.19967.

AI USE STATEMENT

We used generative AI tools, primarily ChatGPT and OpenAI Codex, as assistants at several stages of this work. They were used to support brainstorming and refinement of research hypotheses, provide feedback on experimental design and controlled evaluation settings, assist with the conceptual and theoretical framing of the observed phenomenon, and help interpret experimental results. Generative AI tools were also used for literature search and summarization, drafting and editing portions of the manuscript, translation and language polishing, generating and refining experimental prompts, assisting with figure design, and writing or editing experimental and visualization code.

All research decisions, experimental configurations, claims, and conclusions were made and verified by the authors. AI-generated methodological suggestions were evaluated through actual experiments rather than accepted without validation. AI-assisted code was inspected and tested by the authors, and cited literature and factual claims were checked against their original sources. The authors reviewed and revised all AI-assisted text and artifacts and take full responsibility for the final content of this work.

A Additional Validation Details and Results

A.1 Tasks and Evaluation Setup

The quantitative control tasks in Section 3 cover angles, ellipse axis-length differences, and progress bar proportions. The ellipse task controls the difference in length between the vertical and horizontal axes, while the progress bar task controls the fraction of the bar that is filled. Progress bar control uses a separately constructed test set.

For the geometric tasks, we construct 10 prompt templates at each position, using five for training and holding out five for testing. Table 5 separates generalization across target values and templates. Unseen target values lie within the range of training values, corresponding to interpolation generalization. The main cross-model comparison uses Different/Out, where both target values and templates are unseen.

Table 5: Evaluation splits for generalization across target values and prompt templates.

Evaluation split	Target values	Prompt templates
Same/In	Seen	Seen
Same/Out	Seen	Unseen
Different/In	Unseen	Seen
Different/Out	Unseen	Unseen

A.2 Generalization across Target Values and Templates

When evaluated using their respective training formats, the Qwen-Image models trained with the condition at the beginning achieve positive scores on all four splits for both angle and ellipse control. The model trained with the angle condition at the end yields negative scores on all four splits. For ellipse control, the model trained with the condition at the end achieves positive scores on seen target values but negative scores on unseen values, regardless of whether the template is seen. The Front and Back rows of Table 3 report the four-split angle scores at the default training budget.

These results show that the angle gap persists across target-value and template splits, while the ellipse task reveals a difference in target-value generalization between training positions.

A.3 Training Duration

We hold the data and optimization configuration fixed for angle control with Qwen-Image and compare training for 4, 8, and 16 epochs. Table 6 reports scores on Different/Out when training and test positions match.

Table 6: Angle control with Qwen-Image under different training durations. Scores are evaluated on Different/Out with matching training and test positions.

Training duration	A_{FF}	A_{BB}
4 epochs	0.1130	−0.1609
8 epochs	0.3339	−0.1986
16 epochs	0.1409	−0.0803

Across all three training durations, front-position training achieves positive scores on all four target-value/template splits when evaluated at its training position, whereas back-position training yields negative scores throughout. The positional gap thus persists across the tested durations.

Extending training from 8 to 16 epochs narrows the gap through both an improvement in the back-position score and a decline in the front-position score. This cannot be interpreted solely as back-position training catching up. Longer training does not eliminate the gap within the range examined, but these results do not establish that back-position training cannot acquire control with a larger budget.

B Additional Mechanistic Analyses and Prefix Probes

B.1 Propagation Probe Setup

For each prompt format in Section 4.1, we replace only the target angle and examine the frozen text encoder's responses. Key influence measures the representation change of the angle token itself, context spread measures the average change across other context tokens, and coverage measures the fraction of context tokens whose responses exceed the probing threshold. Context spread and coverage in Table 2 are measured at layer 28. Generator response is the relative L_2 change in predictions with the image latent, noise level, and noise realization held fixed. The front-position propagation advantage persists across all four target-value/template splits.

B.2 Training Diagnostics

Table 7 collects the available diagnostics from the standard front-position and back-position training trajectories in Section 4.2. Correct-condition ROI loss is compared before SFT and at the end of epoch 8; counterfactual loss gaps are compared at the ends of epochs 1 and 8. Prediction response is measured at the end of epoch 8.

Table 7: Diagnostics along standard Front and Back training trajectories. These measurements are separate from the Prefix comparison in Appendix B.3.

Metric	Checkpoint	Front	Back
Correct-condition ROI loss	Before SFT	0.015815	0.015574
Correct-condition ROI loss	Epoch 8	0.003363	0.003533
ROI counterfactual loss gap	Epoch 1	0.000162	0.000175
ROI counterfactual loss gap	Epoch 8	0.000432	0.000326
Prediction response strength	Epoch 8	0.013759	0.011077

The front-condition counterfactual loss-gap point estimate exceeds the back-condition estimate from epoch 6 onward. At epoch 8, the 95% confidence interval for the front-minus-back difference is $[-0.000007, 0.000239]$, which includes zero. The corresponding interval for the paired prediction-response difference is $[0.000940, 0.004536]$, supporting a stronger response to the front condition. These intervals reflect sampling uncertainty in the paired probe evaluation, rather than variation across independent training runs. The counterfactual loss gap in the background region is close to zero, indicating that condition discrimination is primarily localized to the target angle region.

B.3 Prefix Probes and Selective Perturbations

Mixed-training configuration. The Mixed baseline in Table 3 uses five front-position templates and five back-position templates for 8 epochs and is evaluated at both positions. Its scores remain below the constant baseline under the configuration examined.

Prefix propagation and discrimination. Since the text encoder remains frozen, adding the prefix directly changes the distribution of conditioning representations: at layer 28, Prefix yields a context spread of 0.059246 and coverage of 0.923059, both higher than those of standard back-position prompts.

In an 8-epoch comparison conducted separately from the training-trajectory probes in Section 4.2, standard back-position training and Prefix fine-tuning produce similar ROI losses under the correct condition, at 0.003442 and 0.003394, respectively. Their counterfactual loss gaps are 0.000306 and 0.001211, respectively, with the latter approximately 3.96 times the former. The 95% confidence interval for the paired increase is $[0.000395, 0.001505]$, reflecting sampling uncertainty in the paired probe evaluation. Prediction response strength also increases from 0.010905 to 0.024802. This comparison shows that the improvement in control with Prefix is accompanied by stronger condition discrimination and prediction responses.

The comparison above changes both the trained model and the input format. To further localize the change in condition utilization, we hold each trained checkpoint fixed and, in prompts contain-

ing angle information at both the beginning and end, replace the angle in the prefix, the suffix, or both, and measure the counterfactual loss gap within the ROI (Table 8). Replacing only one occurrence creates diagnostic prompts with inconsistent angle values at the two positions, allowing us to examine the model's dependence on each field.

Table 8: ROI counterfactual loss gaps under selective angle perturbations. Replacing a single occurrence yields an intentionally inconsistent diagnostic prompt.

Condition perturbation	Back-trained model	Prefix-trained model
Replace prefix and suffix	0.000310	0.001211
Replace prefix only	0.000048	0.000897
Replace suffix only	0.000290	0.000375
Remove prefix, then replace suffix	0.000306	0.000259

The model trained with Prefix is more sensitive to prefix replacement: replacing only the prefix yields a larger loss gap than replacing only the suffix. The model trained with standard back-position prompts exhibits the opposite pattern. For the latter model, adding a prefix only at inference changes the loss gap from 0.000306 to 0.000310, without reproducing the discrimination gains obtained through Prefix fine-tuning. For the model trained with Prefix, removing the prefix reduces the loss gap to 0.000259, consistent with the decline in generation control. Together, these results support the interpretation that Prefix fine-tuning increases the model's use of target information at the beginning, and that its benefits are associated with how the model learns to use conditioning information during training.

C Additional Details of the Controlled Shape Experiments

C.1 Training and Evaluation Setup

Controlled environment. The controlled environment contains nine shapes, three colors, and two sizes, yielding $9 \times 3 \times 2 = 54$ canonical rendered images at 32×32 resolution. Stars, hearts, and trapezoids are designated as target concepts; circle, ellipse, square, triangle, cross, and crescent serve as base concepts. Prompts contain three fields describing shape, color, and size.

The model combines a frozen Qwen3-0.6B-Base causal text encoder (Yang et al., 2025a) with a conditional U-Net of approximately 28M parameters and performs diffusion directly in pixel space. The text encoder remains frozen throughout both pretraining and SFT.

Controlled pretraining. We vary target-concept exposure using the scalar α. Each of the three target shapes is sampled with probability $\alpha/9$, so their total probability mass is $\alpha/3$; the remaining probability is distributed uniformly across the six base shapes. Pretraining covers all six permutations of the shape, color, and size fields, runs for 10,000 steps with batch size 256, and uses pretraining seed 0.

The dense sweep uses 16 exposure values:

$$\alpha \in \{0,\ 0.00005,\ 0.0001,\ 0.0003,\ 0.0005,\ 0.001,\ 0.00125,\ 0.002,\\ 0.00225,\ 0.0025,\ 0.00275,\ 0.003,\ 0.00325,\ 0.0035,\ 0.00375,\ 0.004\}.$$

Here, α is an intervention on visual exposure rather than a calibrated measure of prior strength. In particular, $\alpha = 0$ only means that the target shapes are absent from the controlled visual pretraining distribution; the frozen pretrained text encoder may still contain concept-related linguistic representations.

Paired SFT. Fine-tuning uses the 18 target combinations formed by three target shapes, three colors, and two sizes. Front prompts place the shape field first, whereas Back prompts place it last. For each α and SFT seed, the paired Front and Back runs share the same pretrained checkpoint, image sequence, diffusion noise, timestep sampling, optimizer initialization, and training configuration, differing only in condition placement.

SFT uses learning rate 10^{-5}, batch size 256, and a budget of 100 steps. We run three SFT seeds, $\{0, 1, 2\}$, and evaluate the following 16 checkpoints:

$$k \in \{0, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 75, 80, 90, 100\}.$$

Step 0 denotes the pretrained checkpoint before SFT.

Generation evaluation. The dense sweep evaluates matched-position control only: Front-trained models are evaluated with Front prompts and Back-trained models with Back prompts. Each checkpoint is sampled with 50-step DDIM. Evaluation covers all 18 target shape–color–size combinations, with four generation seeds per combination, for a total of 72 generated images per model and checkpoint. Shape accuracy is measured using the contour-matching evaluator.

The four generation seeds quantify sampling variability for a fixed trained model. The three SFT seeds instead quantify variability across independently fine-tuned models; these two sources of randomness are treated separately.

C.2 Gap and Uncertainty Measures

Let $G^{(r)}_{FF}(\alpha, k)$ and $G^{(r)}_{BB}(\alpha, k)$ denote matched-position shape accuracies for SFT seed $r \in \{0, 1, 2\}$. The main heatmap reports

$$\Delta_{\text{mean}}(\alpha, k) = \frac{1}{3}\sum_{r=0}^{2} G^{(r)}_{FF}(\alpha, k) - \frac{1}{3}\sum_{r=0}^{2} G^{(r)}_{BB}(\alpha, k). \tag{5}$$

To characterize seed variability conservatively, we additionally compute

$$\Delta_{\text{lower}}(\alpha, k) = \min_r G_{FF}^{(r)}(\alpha, k) - \max_r G_{BB}^{(r)}(\alpha, k), \tag{6}$$

$$\Delta_{\text{upper}}(\alpha, k) = \max_r G_{FF}^{(r)}(\alpha, k) - \min_r G_{BB}^{(r)}(\alpha, k). \tag{7}$$

The interval

$$[\Delta_{\text{lower}}, \Delta_{\text{upper}}]$$

is an observed worst-case range across the three SFT seeds, not a confidence interval or standard deviation.

The gray and purple contours in Figure 2 are computed from

$$G_{FF}^{\text{mean}}(\alpha, k) = \frac{1}{3} \sum_{r=0}^{2} G_{FF}^{(r)}(\alpha, k).$$

They mark regions with $G_{FF}^{\text{mean}} < 15\%$ and $G_{FF}^{\text{mean}} > 90\%$, respectively. These thresholds are used only to visualize low-learning and near-saturated regions and are not interpreted as calibrated phase boundaries.

Several points illustrate the robustness of the intermediate-regime effect. At $\alpha = 0.002$, the mean Front advantage is 16.2 percentage points at step 10, with an observed range of $[6.9, 22.2]$, and 19.4 points at step 20, with range $[8.3, 30.6]$. At $\alpha = 0.003$, the gap reaches 23.6 points at step 25 with range $[19.4, 30.6]$, remains 21.3 points at step 35 with range $[13.9, 29.2]$, and 20.8 points at step 40 with range $[13.9, 30.6]$. In contrast, once matched-position accuracy approaches saturation, the observable gap becomes small.

C.3 Discriminative and Generative Support

The controlled exposure parameter α, discriminative support, and observable generation accuracy describe different quantities. In an auxiliary probe at $\alpha = 0.0001$, we construct counterfactual conditions among the three target shapes and compare native prediction losses. Pairwise discrimination reaches 62.7% even though direct target-shape generation accuracy before SFT is zero.

This shows that failure to generate a target shape reliably does not imply complete absence of condition-sensitive information. Conversely, discriminative preference for the correct condition does not guarantee usable generation control. The auxiliary probe uses three target-shape candidates and is therefore not directly comparable in scale to the eight-candidate pre-adaptation probe used for the large-model experiments in Section 5.1.