CIRCUIT MECHANISMS FOR SPATIAL RELATION GENERATION IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although Diffusion Transformers (DiTs) have greatly advanced text-to-image generation, models still struggle to generate the correct spatial relations between objects as specified in the text prompt. Although mechanistic interpretability studies have been adopted to explain neural networks' behavior in language and vision transformers from the perspective of the internal computation of representations, they have not yet been used to study how a DiT can generate correct spatial relations between objects. In this study, we investigate this open problem in a controlled setting. We train, from scratch, DiTs of different sizes with different text encoders to learn to generate images containing two objects whose attributes and spatial relations are specified in the text prompt. We find that, although all the models can learn this task to near-perfect accuracy, the underlying mechanisms differ drastically depending on the text encoder. When using random text embeddings, we find that the spatial-relation information is passed to image tokens through a two-stage circuit, involving two cross-attention heads that separately read the spatial relation and single-object attributes in the text prompt. When using a pretrained text encoder (T5), we find that the DiT uses a different circuit that leverages information fusion in the text tokens, reading spatial-relation and single-object information together from a single text token. We further show that, although the in-domain performance is similar for the two settings, their robustness to out-of-domain perturbations differs, potentially suggesting the difficulty of generating correct relations in real-world scenarios.

1 INTRODUCTION

Diffusion and flow model (Sohl-Dickstein et al., 2015; Dhariwal & Nichol, 2021; Ho et al., 2020; Lipman et al., 2023; Albergo et al., 2023) has been leading the charge in generative modeling in many domains, image, video, shape (Ho et al., 2022), etc. Specifically, conditional diffusion transformers (DiT) for text-to-image generation (T2I) have unleashed enormous creativity in both industry and the research community, enabling high-fidelity, diverse image synthesis from natural language prompts Rombach et al. (2022b). However, many current T2I models often fail to follow prompts when composing multiple objects onto a scene (Conwell & Ullman, 2022), particularly in arranging their spatial relations (Huang et al. (2023), Ghosh et al. (2023), Huang et al. (2025)). While the field is fast advancing in generating accurate attributes for single objects, the improvement of generating correct relations between objects is slow (Fig.9). Increasing attention has been drawn to this problem, and multiple remedies have been proposed recently, including layout conditioning, cross attention guidance, curriculum learning and finetuning with domain-specific data (Li et al. (2023), Chefer et al. (2023), Chatterjee et al. (2024), Han et al. (2025)). However, few work has approached this problem by understanding the underlying circuit for correct composition of multiple objects. Inspired by an emerging research field, named "mechanistic interpretability", that reverse-engineer a model's internal computations to identify how neurons, attention heads, and weights implement algorithms and produce specific outputs, we study this relation generation problem in a mechanistic fashion, with a goal to understand how T2I models can generate spatial relations under different configs, and under what conditions they could fail.

To study this problem in a controlled setting, we construct a text-to-image task and train T2I models from scratch. The task is to generate two objects (chosen from 2 colors and 3 shapes) in the scene with one of eight spatial relations specified in the text. Then we delve into the underlying transformer

circuits to achieve this task, and find the actual circuits used to solve this task heavily depend on the choice of text encoders. With random token embedding, the T2I model implements a two-stage circuits with two specialized cross-attention heads for reading relation information and single object information respectively. With T5 text encoder, because the information of the words in the prompt is fused, each token contains the full information of the sentence. We find the T2I with T5 encoder indeed reads all the relation and single-object information from a single token. We justify the circuits we found by both ablation and causal manipulation. We further find that though the two circuits mechanism achieves similar task accuracy, their robustness is different upon small perturbation in the text prompt. The accuracy with T5 encoder collapses after perturbing by adding an extra token in the prompt.

Our study resolves several open questions: 1) It was unclear how neural networks encode and use non-commutative relations between objects (Wattenberg & Viégas, 2024). Our work reveals a concrete circuit in diffusion transformers that image tokens can read and implement the relational information in the text, offering a mechanistic example that may generalize to other relational reasoning tasks. 2) The iterative nature of sampling has been an obstacle that complicates attention map analysis and circuit finding. We provide a systematic approach to summarize attention maps and pinpoint heads underlying certain communication patterns, which could be adapted as a general tool to study DiT. 3) Previous studies attributed the spatial relation generation failure to particular stages, e.g. cross attention (Chefer et al. (2023), Phung et al. (2023)) or text encoding (Zhang et al. (2024), Kang et al. (2025)). In this study, we offer a holistic view that bridges these threads. In our toy setting, the T5-based DiT relies on the information fusion by T5 for spatial relationships while the RTE-based DiT implements its own circuits for generating relations. This suggests that the embedding model could be the bottleneck for generating spatial relations in real-world scenarios, making embedding model improvements more critical than DiT modifications.

2 Background

Spatial generation failure in T2I models Failure in multi-object spatial relation generation has been widely reported in T2I models. While new models improved significantly in generating accurate single object attributes, the improvement of generating correct relations between objects is mild (Fig.9). One common view is that spatially localized cross-attention grounds object placement. Building on this hypothesis, recent work tackles spatial-relation failures by directly manipulating attention at inference. Attend-and-Excite Chefer et al. (2023) "excites" the cross-attention to subject tokens to prevent catastrophic neglect and improve attribute binding; and Grounded T2I with Attention Refocusing Phung et al. (2023) optimizes cross- and self-attention using layout-derived losses—via user boxes or LLM-proposed layouts—to enforce multi-object placement and spatial relations. On the other hand, works like Zhang et al. (2024) and Kang et al. (2025) argue that the poor spatial performance of T2I models stems from the limitations of text encoders. Zhang et al. (2024) finds that text encoders used across frontier T2I models do not sufficiently preserve spatial relations information in their encodings. Similarly, Kang et al. (2025) argues that multiple image properties including spatial relations cannot be simultaneously encoded by CLIP.

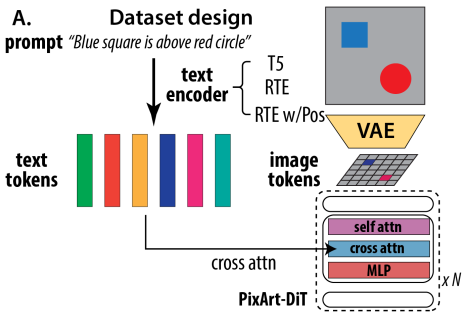


Figure 1: **Schematics of the model and task**. The T2I model archetecture adoptes the design of PixArt. There are three main components: the text encoder that preprocess text tokens, the DiT which is the backbone of the image diffusion denosing process, and the cross attention that passes text information to image tokens. The task is to generate two objects with a specified spatial relation.

Interpretability in diffusion The work that shares the most similar interest with us is (Okawa et al., 2024; Park et al., 2024), which studied the learning dynamics of composition of attributes on a single object in a conditional diffusion model using a minimalist dataset. The key difference is that we focus on the generation of a composition of multiple objects on the scene instead of a single

object. Another difference is that we study the architecture where the conditioning signal is encoded by a set of word vectors (e.g. T5 encoder) and passed the information to image tokens via cross attention, instead of a single vector summing all attributes as in Okawa et al. (2024). This setting is more closely related to modern text2image frameworks (Chen et al., 2023; Rombach et al., 2022a; Xie et al., 2024).

3 TRAIN T2I MODELS TO GENERATE SPATIAL RELATION

Previous work demonstrates that pretrained T2I models across different text encoders and architectures show disproportionally better capabilities for single object feature generations than object 2D (top-bottom, left-right) or 3D (front-back) relations (Huang et al., 2023). This observation motivates us to study the mechanism of spatial relation generation and why it frequently fails in T2I models. To understand this problem in a controlled manner, we construct a minimal text-image dataset and train DiT-based T2I models of different sizes and text encoders from scratch. We make sure that both single object features and object relation properties are reliably learned and amenable to mechanistic analysis.

Dataset setup We reason that such a dataset should have following properties: 1) multiple objects in the scene with distinct features, 2) objects arranged to satisfy specific (spatial) relations described by the prompt, and 3) samples simple enough to evaluate rigorously. Guided by this principle, we design a dataset of the following format: each sample consists of a pormpt in the format `[descriptor A] [object A] [relation] [descriptor B] [object B]`", e.g., *red square above and to the left of blue circle*", and a corresponding image with two objects positioned on a gray background (Fig. 1**A**). We use three shapes (*circle, triangle, square*), two colors (*red, blue*), and eight spatial relations: *left, right, above, below, upper left, upper right, lower left, lower right*. The shape and color of the two objects are always distinct, and their positions are arranged to avoid collisions. The color descriptors A and B are randomly dropped, and spatial relations are described with multiple paraphrases to add variability.

Model architecture We use a DiT-based T2I model, following a PixArt-style architecture representative of the state-of-the-art open-source DiT models (Chen et al. (2023)). We train several model sizes with patch size 2: DiT-B (12 layers, 12 heads, 768 latent dimensions), mini (6L, 6H, 384-d), micro (6L, 3H, 192-d), and nano (3L, 3H, 192-d). Following common practice, images are encoded with the Stable Diffusion VAE (Rombach et al., 2022a). As for text conditioning, we compare three encoders: (i) T5-XXL (Raffel et al., 2023); (ii) a random token encoder with sinusoidal positional encoding (RTE) (iii) RTE without positional encoding. This comparison tests whether the diffusion transformer can learn object relations without semantic or contextual structure in the text embeddings, enabling better localization of the relational computations (Fig. 1**A**).

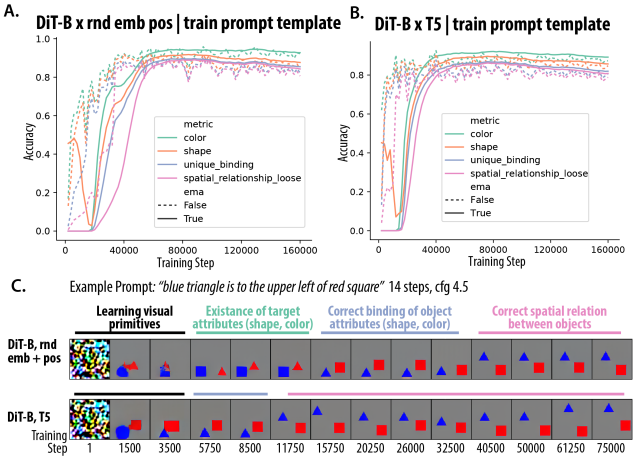


Figure 2: **Training dynamics of the T2I models (DiT-B)**. **A)** and **B)** Both models trained with random token embedding and T5 can achieve good accuracy on the task. **C)** The task is learned step by step. In both models, they first learn to generate objects but with wrong attributes binding, then they the correct binding of single-object attributes (e.g. red square), finally they learn the correct spatial relation.

Training dynamics Throughout training, we evaluate model generations on 96 prompts spanning 8 spatial relations and 12 object pairs. Generation is performed on multiple random seeds with the default sampler (DPM-Solver++ with 14 steps (Lu et al., 2022)) and classifier-free guidance of 4.5 (Ho & Salimans, 2022). We

evaluate the consistency of generated images with the prompts using classic segmentation and classification tools from `cv2` (Bradski (2000)). Specifically, we assess on the following 4 aspects of features: 1) existence of correct colors on the image, denoted as `color`, 2) existence of correct shapes on the image, denoted as `shape` 3) correctness of shape color binding on the two objects, denoted as `unique_binding` and 4) correctness of spatial relation between the identified two objects, denoted as `spatial_relation`.

Accuracy on all four metrics increases with model parameter sizes up to the DiT-mini configuration; accuracy gains from DiT-mini to DiT-B are marginal. All trained models at the largest parameter size (DiT-B) show high accuracy in `color` and `shape`, but the `unique_binding` and`spatial_relation` accuracy varies significantly depending on the chosen text encoders (Tab. 2). RTE and T5 achieve strong `unique_binding` and `spatial_relation` accuracy, whereas RTE without positional encoding is significantly worse on these metrics. Without positional cues, "red A on top of blue B" and "blue B on top of red A" collapse to the same bag-of-words embeddings, yielding identical outputs. Adding positional information resolves this ambiguity, indicating that pretrained semantic structure (T5) is not strictly required for learning object relations

Having established the end-point performance trends across models, we next examine how these capabilities emerge during training. For both T5-DiT and RTE-DiT models, we evaluate an exponential moving average (EMA) of weights (Karras et al., 2024) following the diffusion-model practice and show the accuracy curves averaged across multiple runs. We consistently observe that `color` accuracy converges first, followed by `shape` and then `unique_binding`. `spatial_relation` is learned the slowest (Fig. 2**A.**), indicating that relational composition is more challenging to learn than single-object attributes or bindings. Comparing across the two text encoders, we observe that T5-DiT models converge to optimal accuracy faster across all 4 features. Moreover, the temporal gaps between different feature learning are tighter. We also provide generation examples at different checkpoint steps throughout the training for visual examination. The different dynamics suggests that the two family of models potentially use different internal mechanisms to accomplish the same generation task.

4 Relation generation circuits in RTE-DiT

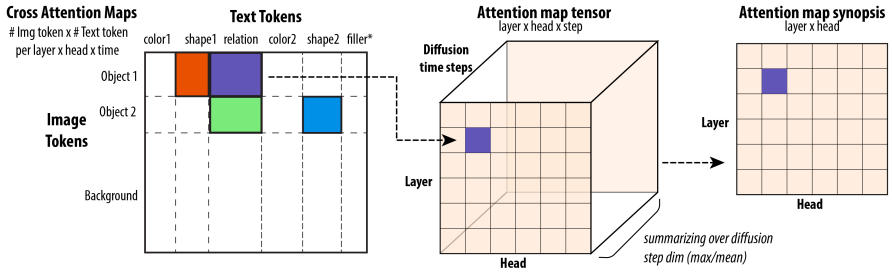


Figure 3: **Illustration of our method to find relevant heads**, which we name as "attention synopsis". The giant attention tensor is first reduced to those only between two interested groups of tokens (e.g. the relation token regardless specific words, or an object token regardless where or what it is). Then the reduced attention tensor is averaged over diffusion time steps, resulting in a layer $\times$ head map which we use to pinpoint relevant heads.

4.1 Attention synopsis

In our DiT-based model architecture, cross-attention mechanism is the sole pathway for text prompts to influence the image tokens at each denoising step. Therefore, we examine the cross-attention patterns to gain insights into how text on single object feature and spatial relations guide correct generations. Given the high dimension of DiT's cross attention maps ([layers $\times$ heads $\times$ time steps $\times$ condition vs uncondition pass $\times$ number of tokens]), it is impractical to perform manual inspection. Moreover, simply averaging attention maps over different samples and prompts can obscure specific interactions. Therefore, we develop a scalable paradigm to analyze and quantitatively summarize the cross-attention head patterns called *Attention Synopsis* (Fig. 3) Leveraging this method, we efficiently search through over 10 million attention maps to trace text-to-image flow and localize the relevant

circuit mechanisms for spatial relations generation. Specifically, we leverage the fact that token categories are identifiable in both image and text (image tokens by object segmentation, text tokens by semantic attribute). We then aggregate attention within and across categories, yielding interpretable category-to-category interaction patterns. After this aggregation, we reduce the cross-attention map tensor dimension to [num layer $\times$ num head $\times$ num time steps]. Given that the attention maps usually change smoothly across time, we further calculate the mean attention maps over time steps, reducing the tensor to shape `[num layer, num head]`, which we denote as the *attention map synopsis*.

Many previous works have reported the cross attention communication between the text token of a single object and the corresponding object in the image (Tang et al., 2022), and this property is leveraged to control generation (Hertz et al., 2022; Liu et al., 2024). These findings suggest that there is coupling between single object tokens across text and image modality, supporting good generation. However, less is known about whether text tokens describing spatial relations *between* objects extends similar properties. Therefore, we leverage the *Attention Synopsis* method to examine all category-to-category cross-attention patterns, especially focusing on the spatial relations category. We show results for RTE-DiT in this section and T5-DiT in Section 5.

In RTE-DiT, we find a minimal circuit that enables generation of correct objects at correct spatial locations. The circuit consists of two key cross attention heads which we discuss in details below.

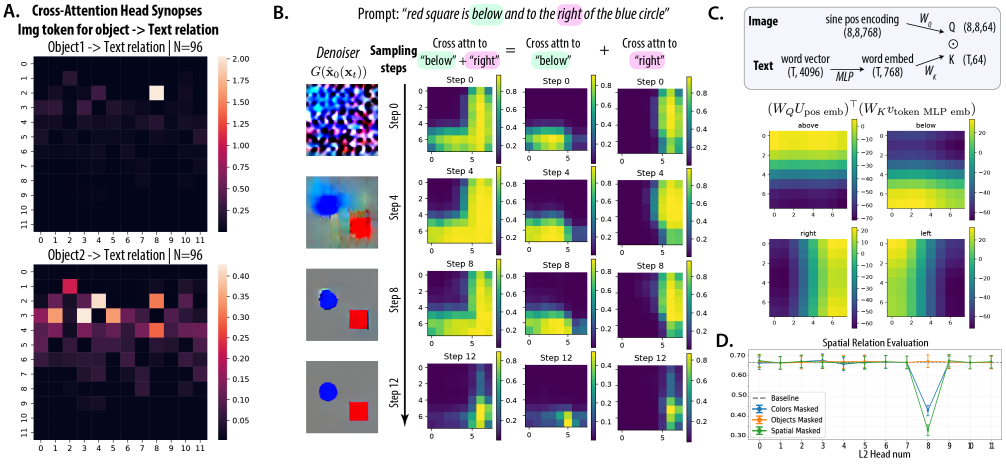


Figure 4: **The spatial relation heads in random-embedding-based T2I**. **A)** We find specialized cross attention heads that contributes to the object image tokens (top: the object1 in the text; bottom: the object2 in the text) attending to the relation text tokens. **B)** We show the activation of this head across images tokens and sampling steps. The map for the composite relation "below and right" decomposes cleanly as the sum of the maps for "below" and "right", **C)** The observed attention patterns can be induced by positional embedding.

4.2 Spatial Relation Head

We first use attention synopses to look for cross attention heads that contribute to passing relation information from text tokens to image tokens and find specialized heads (Fig.4). We find, indeed, there are specialized heads for this job. In the case of "object1" (i.e. the first object in the text), there is only one head (L2H8) that dominates, while in the case of "object2", there are a small number of heads that have this role. We also find similar pattern in models with other sizes.We name these heads "spatial relation head". By plotting the "activation" of this head across sampling steps (Fig. 4), we find its spatial pattern aligns with the spatial regions corresponding to the relation token, with a trend of concentrating to the object. This observation implies the query that maximizes the attention score of this head is potentially aligned with positional embedding. We verify this by checking the QK circuit Elhage et al. (2021) of L2H8 (Fig.4C). The head projects sinusoidal positional embeddings from image tokens into the query space (Q) and MLP-projected relation-word embeddings[1] into the

[1]In the PixArt architecture, frozen text embeddings are first passed through a learnable MLP projection before entering the attention layers.

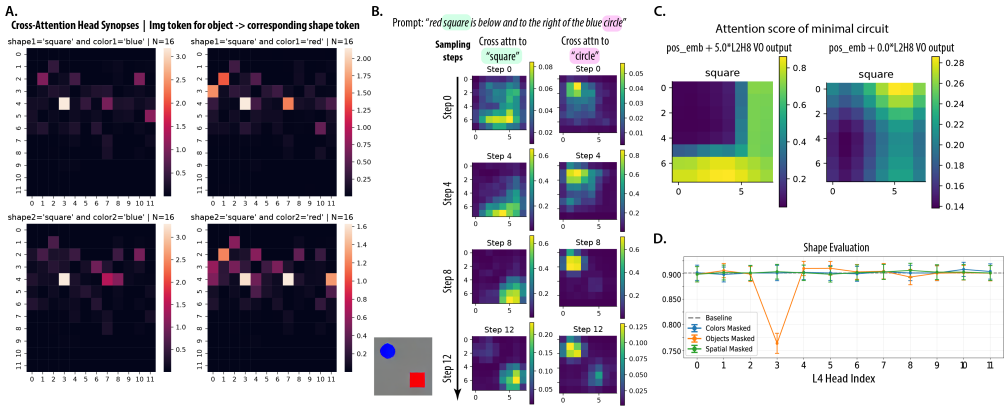
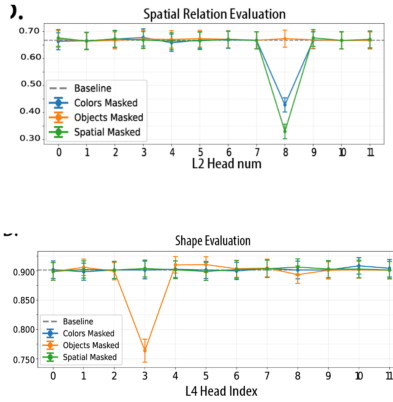


Figure 5: **The object generation heads in random-embedding-based T2I**. **A)** We find specialized heads in the synopses of cross-attention, computed from image tokens of each object to its own shape tokens. **B)** We show the activation of this head across images tokens and sampling steps for the prompt "red square is below and to the right of the blue circle": tokens at the eventual square location attend to "square," while the other object attends to "circle"; selectivity sharpens from Step 0→12. **C)** Injecting the VO output of the relation head (L2H8) into positional embeddings is sufficient to elicit selective attention from tagged locations to the "square" token (left); without the tag the pattern is weak (right). This indicates the object generation head reads the relational tag generated by the spatial relation head.



For Binxu, Jingxuan
有一版新的图，我们
去替换这个图

Figure 6: **Attention head ablation reveals focused effect**.

key space (K) via learned linear layers (W_q, W_k). This QK interaction aligns specific coordinates in the image grid with the semantics of spatial relation tokens. The resulting inner-product maps (Fig.4**C**) form smooth gradients whose orientation reflects the spatial relations (e.g., "above" produces a vertical gradient). These gradients act as positional tags, marking the regions of the canvas where the first object should be placed. Downstream heads then read these tags to guide accurate object placement and generation.

4.3 Object Generation Head

The spatial relation head allows differential tagging of image tokens based on relational text tokens. To successfully complete the task, the model also needs to generate the correct object on the tagged canvas. To this purpose, we examine cross attention heads that contribute to passing object shape information from text tokens to image tokens. We identify a single head—Layer 4, Head 3 (L4H3) that consistently mediates communication between an object's image tokens and its corresponding shape word in the prompt (Fig. 5**A**). This linkage is invariant to both the object's position in the sentence and the specified spatial relation, indicating that the head encodes shape identity independently of relational context. During sampling, we can see this communication channel is active later in sampling (step4-8), linking each objects to their corresponding shape tokens in the prompt (Fig. 5**B**).

4.4 Ablation of Casual Manipulation

To test whether the above discussed heads has a causal role in correct spatial relation and object shape generation, we perform both ablation and casual manipulation experiments.

We also perform layer- and head-specific ablation in image to text tokens cross-attention and evaluate image generation on all testing prompts. Specifically, we identify and mask tokens corresponding to 3 types of concepts in the text prompts respectively - object shape (`object`), object color (`color`) and spatial relation (`spatial`). Ablating spatial-relation attention specifically in L2H8 reduced relational accuracy from 67% to 33%, while other heads showed negligible effects (Fig. 4**D**). This confirms L2H8's critical role in implementing the correct spatial layout. On the other hand, ablating `shape` cross-attention particularly in L4H3 makes object shape generation accuracy decrease from 90% to 76% while ablation of other layer head combinations shows minimal effect (Fig. 6**B**). This emphasized the critical value of such a head in robustly generating correct object shape. Although the effect size of object shape ablation is smaller than the spatial relations ablation, effects in both cases are only confined to the two previously identified heads, L2H8 and L4H3, suggesting a highly concentrated circuit.

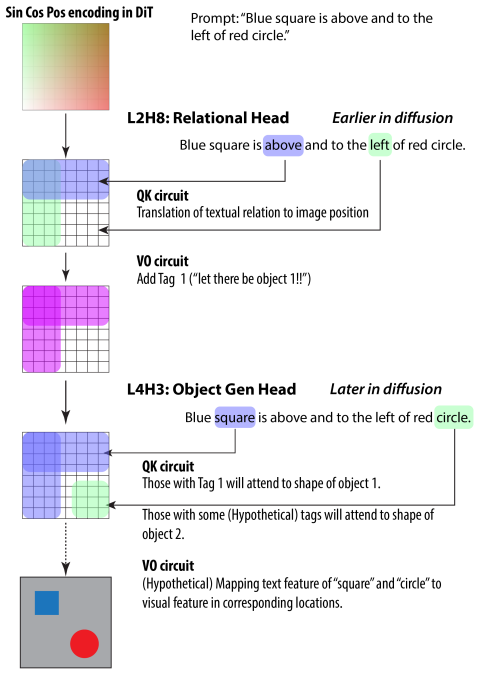


Figure 7: Schematics of the object relation circuit in DiT trained with random embedding

We reason that L2H8 and L4H3 functions in sequence to generate a correct object at the correct spatial location. To test this hypothesis, we inject the VO output of the relation head (L2H8) into L4H3's image tokens positional embeddings. This manipulation is sufficient to elicit selective attention from tagged locations to the "square" token (Fig.5**C.**). Without the injected VO inputs, no obvious attention pattern is observed (right), showing that the downstream object shape head reads the relational tag.

4.5 CONSISTENCY ACROSS MODEL SIZES

Finally, we test the generalizability of this circuit mechanism by examining across RTE-DiT of different parameter sizes. We find putative spatial relation heads consistently in DiT models of three different scales (DiT-B, mini, micro) we trained. The smallest model (DiT-nano) where we cannot find such a head also failed badly on spatial relation (accuracy 5%) (Tab. 2).

In summary, for DiTs trained with random token embedding (RTE), relational object generation unfolds in two stages (Fig. 7): The **"spatial relation head"** reads relational text tokens (e.g., "above," "left") via the QK circuit and interact with the sinusoidal positional encoding of image tokens, producing spatial gradients for each relation. The VO circuit writes positional tags (e.g., Tag 1) onto image tokens, marking where the object (e.g. 1st in sentence) should appear. In the **"object generation head,"** tokens with matching tags attend to shape token of the corresponding object. The VO circuit maps these text features (e.g., "square", "circle") into visual features at tagged locations, generating the object via denoising.

This modularizes operation where relation heads laying the ground and object heads assigning attributes provides a clean and disentangled mechanism for robust composition of relation and object combination.

5 RELATION CIRCUITS IN T5

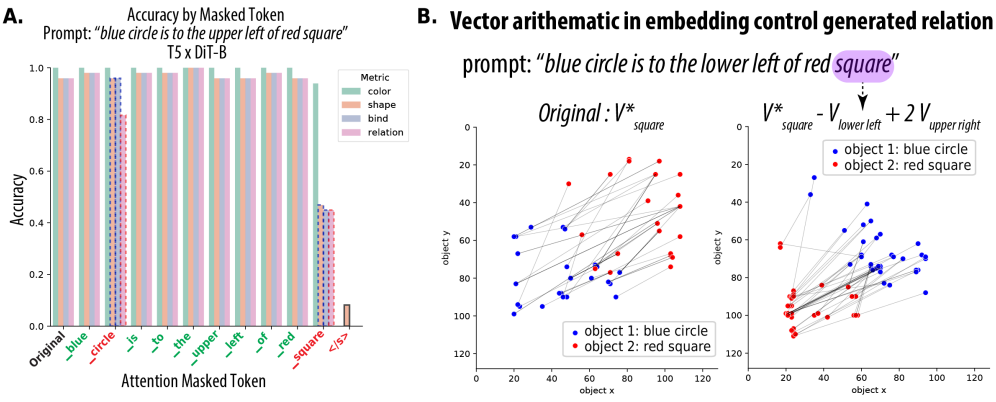


Figure 8: T5 mechanism for relational generation. **A.** T5-based DiT is robust to attention ablation of relation word, but most sensitive to shape2 and EOS. **B.** Manipulation via factorized word vector arithmetic causally affects generated object relation.

Given the clear mechanism found in RTE-based T2I models, it's tempting to apply the same *Attention Synopsis* method and identify specialized cross attention heads for spatial relation in T5-based models. However, no clear pattern emerges from averaging.Thus, we seek alternative strategy for the alternative mechanism.

Given that after T5 encoder each text token could contain the information of the whole sentence, the image tokens may receive the spatial relation information from non-relation tokens. We test this by using attention mask to see which word has the largest effect on object relation generation (Fig.8**A**). Surprisingly, the relation words, filler and color words have little effect on the generation performance (Fig. 8**A**). Ablating `<end_of_sentence>` token disrupts the denoising generation process, thus decimating all evaluation metrics. The relation accuracy decreases after masking `shape1` and `shape2`, where ablating `shape1` reduces spatial relation by 15%, while ablating `shape2` harms all shape, binding and relation accuracy by 50%.

This suggests an intriguing alternative mechanism: through the T5 language model, the information of object1 and object2 and their relation has been encoded in the contextual embedding of `shape2` and `shape1` tokens, thus DiT only needs to decode the information from them, and ignore others.

Visualization (UMAP, tSNE, PCA) of the T5 contextual embedding for the second shape token (V^*_{shape2}) shows clear qualitative separation by spatial relation and, within each relation, by object2's

color. This indicates that multiple semantic factors—its own identity, the other object's identity, and the spatial relation—are jointly represented in this embedding space (Fig. A.1).

We further analyze this effect by a formal variance analysis. We model the contextual embedding of the second object token as a linear combination of four factor vectors: $V^*_{\text{shape2}} = V_{\text{shape2}} + V_{\text{color2}} + V_{\text{shape1}} + V_{\text{rel}}$. This structure allows us to extract a vector for each level within a factor (e.g., a vector for "upper right" versus "lower left"), enabling controlled embedding manipulations. Variance partitioning supports this factorization (Tab. 5). In the raw T5 embedding, `shape2` accounts for the largest share of variance ($\sim$37.5% partial R^2), with `relation` still contributing substantially ($\sim$12%). After projection through the DiT MLP, the balance shifts: `relation` becomes the dominant factor ($\sim$21%), while `shape2`'s share decreases, suggesting DiTs reorganize the token representation and accentuate the relation information for generation.

		V_{shape2}	V_{color2}	V_{shape1}	V_{rel}	tot. R^2
T5 emb	part. R^2	37.5%	4.7%	5.0%	12.1%	73.2%
	marg. R^2	51.4%	4.7%	18.9%	12.1%	
DiT MLP	part. R^2	14.9%	8.0%	7.2%	21.3%	53.4%
	marg. R^2	16.9%	8.0%	9.0%	21.3%	

Table 1: Variance partitioning of T5 embedding and DiT-MLP projection of `shape2` token.

To causally test the hypothesis that relation information is encoded in specific object tokens within the T5 embedding, we performed targeted vector arithmetic on the 4096-d prompt embedding for a shape token (e.g., "square"). Starting from the original embedding of object2, (e.g. V^*_{square} in prompt "blue circle is to the lower left of red square"), we subtracted the learned factor vector for the original relation (e.g., $V_{\text{lower left}}$) and added a scaled vector for an alternative relation (e.g., $2V_{\text{lower right}}$ or $2V_{\text{upper right}}$).

As shown in Fig. 8**B**, this manipulation systematically shifts the generated object positions to match the new relational configuration, while leaving object identities (color and shape) largely intact. This provides direct causal evidence that relational geometry is embedded in the contextual representation of the `shape2` token, and that simple linear operations in this space can reconfigure spatial relations in generated images.

Robustness of RTE-based and T5-based circuits When evaluating on prompt with the exact format as the training ones, RTE- and T5-trained models have comparably high performance on spatial relation. However, slight prompt variation breaks the tie, i.e. adding *the* to the prompt reduces the relational accuracy of T5-DiT model by around 40% (Fig. 10**B.**, Tab. 2). This suggests that even though the task accuracies are similar between RET-based and T5-based T2I models, their robustness to the small perturbations in the text is different. The T5-based model is more sensitive to the perturbation.

6 Discussion

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions, November 2023.

Gary Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it right: Improving spatial consistency in text-to-image models, 2024. URL https://arxiv.org/abs/2404.01197.

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, 2023. URL https://arxiv.org/abs/2301.13826.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis, December 2023.

Colin Conwell and Tomer Ullman. Testing Relational Understanding in Text-Guided Image Generation, July 2022.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. https://arxiv.org/abs/2105.05233v4, May 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Xu Han, Linghao Jin, Xiaofeng Liu, and Paul Pu Liang. Progressive compositionality in text-to-image generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=S85PP4xjFD.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion 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. 8633–8646. Curran Associates, Inc., 2022.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–17, May 2025. ISSN 1939-3539. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3531907.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes!, 2025. URL https://arxiv.org/abs/2503.08723.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and Improving the Training Dynamics of Diffusion Models, March 2024.

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, 2023. URL https://arxiv.org/abs/2301.07093.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing, March 2024. URL http://arxiv.org/abs/2403.03431. arXiv:2403.03431 [cs].

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task, February 2024.

Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space, December 2024.

Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing, 2023. URL https://arxiv.org/abs/2306.05427.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022a.

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*, pp. 10684–10695, 2022b.

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, July 2015. PMLR.

Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

Martin Wattenberg and Fernanda B. Viégas. Relational Composition in Neural Networks: A Survey and Call to Action, July 2024.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers, October 2024.

Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. Compass: Enhancing spatial understanding in text-to-image diffusion models, 2024. URL https://arxiv.org/abs/2412.13195.

A Extended Results

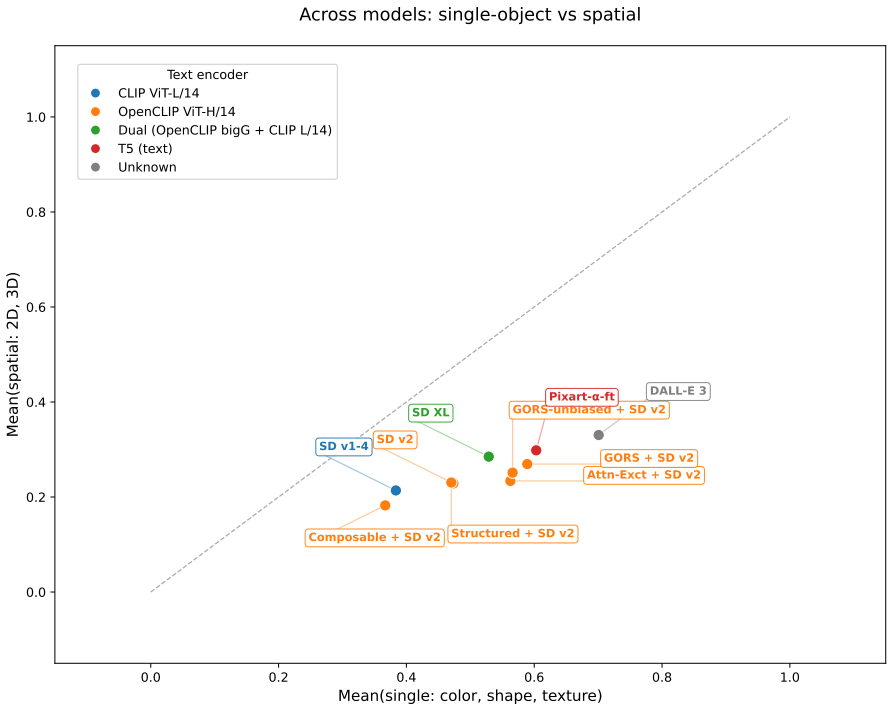


Figure 9: Evaluation of spatial vs single feature accuracy of popular text to image models

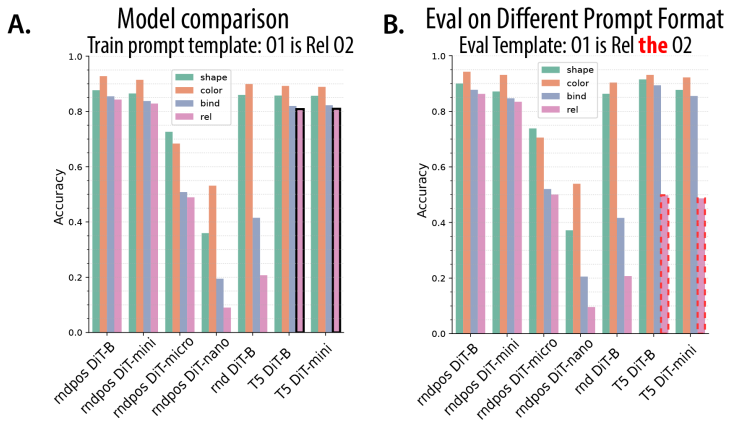


Figure 10: **Evaluation of model performance on trained and generalized prompt template.**

A.1 Evaluation Training Dynamics

Table 2: **Comprehensive evaluation of models on prompt template variations.**
Abbreviations: WD: weight decay, rnd: random embedding, rndpos: random embedding plus position encoding. *bind*: unique and correct attribute binding. *sp rel*: spatial relation correctness (loose). *sp rel+*: spatial relation correctness (stringent). *Dx*, *Dy*: difference of coordinates between the two identified objects (with target attributes) $x_1 - x_2$, $y_1 - y_2$, with the unit pixel (128 pixel total).
All statistics are averaged from 264 prompts, covering all 8 relations and all object combinations, each drawing 50 samples. Thus, the non-zero value in Dx, Dy suggests systematic bias in spatial relation.

model name	template	shape	color	bind	sp rel	sp rel+	Dx	Dy
rndpos DiT-B	O1 is Rel O2	0.877	0.928	0.855	0.843	0.758	-0.4	-0.9
	O1 is Rel the O2	0.900	0.942	0.877	0.862	0.717	-0.3	-0.4
	O1 Rel O2	0.858	0.909	0.833	0.823	0.752	-1.2	-1.6
	O1 Rel the O2	0.877	0.925	0.853	0.842	0.759	-0.3	-1.2
	the O1 is Rel the O2	0.895	0.946	0.868	0.833	0.614	-0.9	0.5
rndpos DiT-mini	O1 is Rel O2	0.865	0.914	0.838	0.828	0.644	0.8	0.4
	O1 is Rel the O2	0.871	0.931	0.847	0.834	0.613	1.1	1.5
	O1 Rel O2	0.778	0.845	0.743	0.737	0.621	1.5	-0.4
	O1 Rel the O2	0.799	0.879	0.770	0.762	0.616	1.8	0.0
	the O1 is Rel the O2	0.767	0.912	0.721	0.680	0.471	-0.1	1.8
rndpos DiT-micro	O1 is Rel O2	0.726	0.683	0.508	0.489	0.315	-0.2	0.2
	O1 is Rel the O2	0.738	0.705	0.520	0.501	0.312	0.3	0.2
	O1 Rel O2	0.626	0.604	0.395	0.386	0.270	0.1	-1.5
	O1 Rel the O2	0.649	0.639	0.410	0.401	0.269	-0.2	-1.2
	the O1 is Rel the O2	0.665	0.724	0.432	0.403	0.234	2.1	-0.7
rndpos DiT-nano	O1 is Rel O2	0.360	0.531	0.195	0.090	0.049	3.2	-0.1
	O1 is Rel the O2	0.372	0.539	0.205	0.096	0.051	2.4	-1.6
	O1 Rel O2	0.270	0.568	0.146	0.069	0.037	5.1	-0.4
	O1 Rel the O2	0.279	0.581	0.151	0.071	0.036	3.2	-2.4
	the O1 is Rel the O2	0.399	0.632	0.193	0.082	0.047	-3.6	-3.5
rnd DiT-B	O1 is Rel O2	0.859	0.899	0.415	0.207	0.192	-0.1	0.1
	O1 is Rel the O2	0.863	0.903	0.416	0.207	0.190	-0.0	-0.0
	O1 Rel O2	0.856	0.893	0.412	0.205	0.191	0.0	-0.0
	O1 Rel the O2	0.860	0.902	0.415	0.206	0.191	-0.0	0.1
	the O1 is Rel the O2	0.866	0.910	0.417	0.207	0.188	0.1	0.0
T5 DiT-B	O1 is Rel O2	0.857	0.892	0.820	0.808	0.749	-0.8	-0.5
	O1 is Rel the O2	0.915	0.931	0.894	0.498	0.306	-33.7	-24.9
	O1 Rel O2	0.853	0.871	0.825	0.608	0.493	-4.5	-16.6
	O1 Rel the O2	0.941	0.958	0.925	0.400	0.217	-35.1	-37.0
	the O1 is Rel the O2	0.917	0.935	0.896	0.529	0.309	-18.4	-20.7
T5 DiT-mini	O1 is Rel O2	0.856	0.889	0.822	0.810	0.659	-0.4	-0.6
	O1 is Rel the O2	0.877	0.922	0.855	0.487	0.259	-35.8	-24.1
	O1 Rel O2	0.816	0.844	0.772	0.559	0.399	-12.7	-18.2
	O1 Rel the O2	0.895	0.946	0.878	0.391	0.184	-38.7	-37.6
	the O1 is Rel the O2	0.906	0.947	0.885	0.537	0.272	-17.1	-19.3
T5 DiT-B WD	O1 is Rel O2	0.183	0.114	0.033	0.033	0.031	-1.2	1.3
	O1 is Rel the O2	0.169	0.104	0.030	0.017	0.013	-39.7	-22.2
	O1 Rel O2	0.164	0.110	0.032	0.025	0.023	-5.0	-15.4
	O1 Rel the O2	0.181	0.122	0.037	0.016	0.011	-40.4	-34.8
	the O1 is Rel the O2	0.160	0.100	0.028	0.017	0.013	-15.0	-18.4
T5 DiT-mini WD	O1 is Rel O2	0.894	0.942	0.866	0.854	0.667	-0.4	1.0
	O1 is Rel the O2	0.911	0.967	0.886	0.521	0.265	-42.0	-19.1
	O1 Rel O2	0.843	0.886	0.804	0.596	0.429	-9.1	-16.7
	O1 Rel the O2	0.911	0.975	0.888	0.414	0.189	-47.1	-35.3
	the O1 is Rel the O2	0.911	0.965	0.887	0.514	0.249	-22.4	-13.8

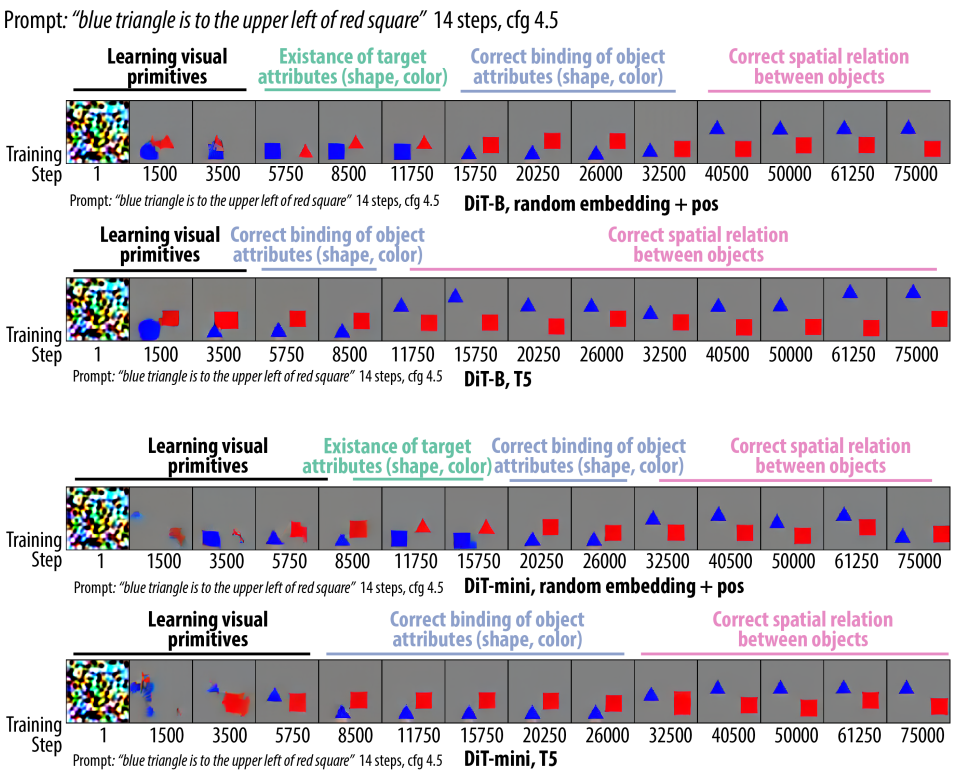


Figure 11: **Comparison of training dynamics of DiT models with different text encoding and scale.** Specific evaluation prompt used was "*blue triangle is to the upper left of red square*", sampled with 14 steps at cfg 4.5, sampled from the same noise seed. Further, T5 models immediately learn to achieve object attribute binding after learning attributes themselves, while random embedding model gradually learn the correct attribute binding and then spatial relation. Across scales, generally, larger scale models train faster.

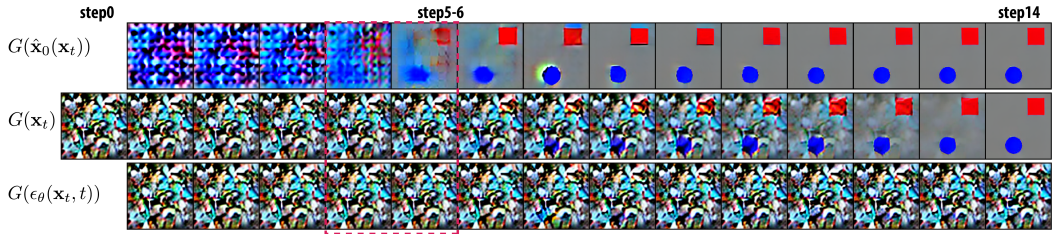


Figure 12: **Observation on sampling dynamics** Specific evaluation prompt used was "*the red square is above and to the right of the blue circle*", sampled with 14 steps at cfg 4.5. Model used is DiT-B rand emb pos. A transition can be seen at step 4-6, where the two object at their final positions can be clearly seen from the expected outcome $G(\hat{\mathbf{x}}_0(\mathbf{x}_t))$.

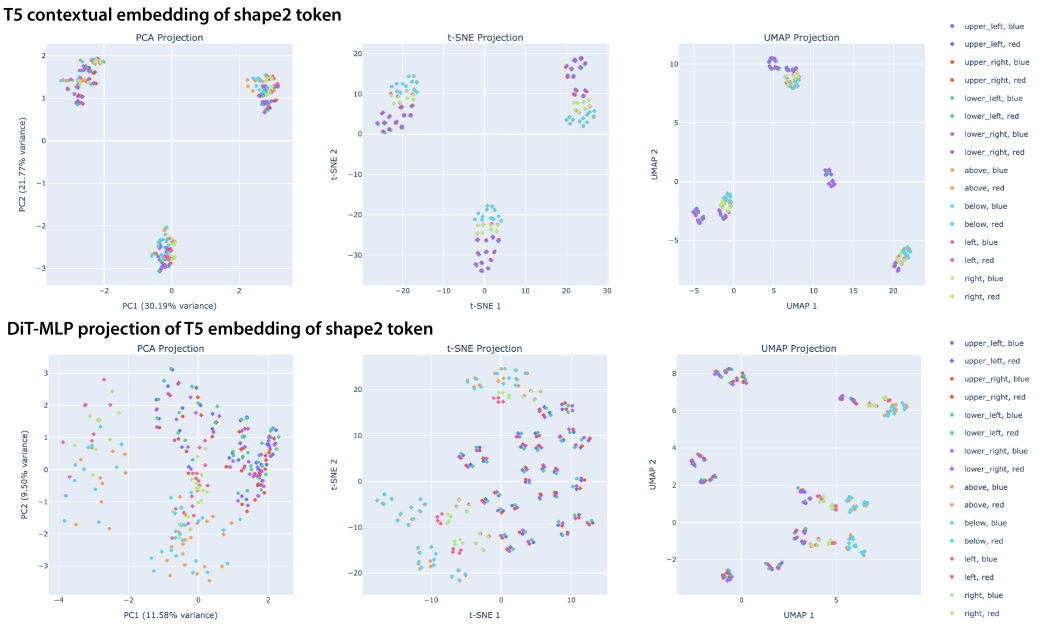


Figure 13: **Dimension reduction visualization of `shape2` token representation (PCA, tSNE, UMAP).** Top row: T5 contextual embedding (4096d), Bottom row: DiT-MLP projection (784d).

B DATASET AND CODE AVAILABILITY

To preserve anonymity, we will publicly release all code, configuration files, and datasets at a permanent URL upon acceptance.

C LLM USAGE

The usage of LLM is limited to language polishing and grammar, and literature search. We asked an LLM to suggest surface-level rewrites to improve clarity, grammar, and style for author-written passages. Edits were limited to phrasing and organization at the sentence/paragraph level. We also used an LLM to source papers, and produce brief literature summaries for writing references.