

Appendix

A Further implementation detail

- Model configuration
- Text prompts for analysis
- Attention perturbation
- Segmentation evaluation metrics
- Further details on image quality metrics

B Additional experiments

- Ablation on timestep choice
- Comparison on I2T and T2I attention maps
- Ablation on normalization method
- Segmentation performance per attention layer and head
- Quantitative results on attention perturbation

C Generalization to other baselines

- Stable Diffusion 3.5
- Flux.1-dev

D Additional visualizations

- Image-to-text (I2T) and text-to-image (T2I) attention map
- Per-head attention maps
- Statistics of attention score and norm
- PCA visualization of attention features
- Emergent behavior of <pad> tokens

E Additional qualitative results for generation

- Extended I2T attention perturbation results
- Extended I2T attention perturbation guidance results
- Image generation results of trained models

F Additional qualitative results for segmentation

- Open-vocabulary semantic segmentation results
- Unsupervised segmentation results

G Limitations of our method

A Further implementation detail

A.1 Model configuration

We utilize the Stable Diffusion 3 (SD3) [2] model for our main analysis. SD3 originally incorporates three vision-language text encoders: CLIP-G/14, CLIP-L/14 [5], and T5-XXL [6]. Due to memory constraints, we disable the T5 encoder and use only the first 77 tokens from the two CLIP encoders. We leverage the attention score from timestep $t = 8$ of 28, where the attention logits are semantically grouped well. Further ablation on timesteps can be found at Sec. B.1. We used classifier-free guidance [4] with scale 7.5 for generation if not specified.

For training and segmentation evaluation, the input images are center-cropped if non-square and resized to a resolution of 1024×1024 . After VAE encoding, the image is further downsampled to 64×64 latent. The resulting attention maps are bilinear upsampled back to the original image size for segmentation evaluation.

A.2 Text prompts for analysis

We borrowed text prompts from DrawBench [7], where diverse categories are included to assess the capability of generative models. From 200 prompts in total, we randomly sampled 50 prompts for our analysis. The full list of selected prompts is shown in Fig. 1.

1. A red colored car.	27. An elephant is behind a tree. You can see the trunk on one side and the back legs on the other.
2. A black colored dog.	28. A pear cut into seven pieces arranged in a ring.
3. A blue colored dog.	29. Rbepfraigerator.
4. A red colored banana.	30. Dining tablez.
5. A white colored sandwich.	31. An instrgument used for cutting cloth, paper, axdz othr thdin mterioal, consamistng of two blades lad one on tvopb of the other and fhasstned in tle mixdqdjle so as to bllow them txo be pened and closed by thumb and fitngesr inserted tgrough rings on kthe end oc thei vatndlzes.
6. A yellow colored giraffe.	32. A bicycle on top of a boat.
7. A green cup and a blue cell phone.	33. A car on the left of a bus.
8. A horse riding an astronaut.	34. Acersecomicke.
9. A shark in the desert.	35. Artophagous.
10. Three cars on the street.	36. Backlotter.
11. One dog on the street.	37. A photo of a confused grizzly bear in calculus class.
12. Two dogs on the street.	38. Photo of an athlete cat explaining it's latest scandal at a press conference to journalists.
13. One cat and one dog sitting on the grass.	39. Hyper-realistic photo of an abandoned industrial site during a storm.
14. Three cats and one dog sitting on the grass.	40. A real life photography of super mario, 8k Ultra HD.
15. Three cats and two dogs sitting on the grass.	41. Colouring page of large cats climbing the eifel tower in a cyberpunk future.
16. A triangular pink stop sign. A pink stop sign in the shape of a triangle.	42. Photo of a mega Lego space station inside a kid's bedroom.
17. An illustration of a small green elephant standing behind a large red mouse.	43. A spider with a moustache bidding an equally gentlemanly grasshopper a good day during his walk to work.
18. A small blue book sitting on a large red book.	44. A bridge connecting Europe and North America on the Atlantic Ocean, bird's eye view.
19. A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.	45. A magnifying glass over a page of a 1950s batman comic.
20. A stack of 3 books. A green book is on the top, sitting on a red book. The red book is in the middle, sitting on a blue book. The blue book is on the bottom.	46. A realistic photo of a Pomeranian dressed up like a 1980s professional wrestler with neon green and neon orange face paint and bright green wrestling tights with bright orange boots.
21. A small vessel propelled on water by oars, sails, or an engine.	47. A sign that says 'Hello World'.
22. A large plant-eating domesticated mammal with solid hoofs and a flowing mane and tail, used for riding, racing, and to carry and pull loads.	48. A sign that says 'Diffusion'.
23. An American multinational technology company that focuses on artificial intelligence, search engine, online advertising, cloud computing, computer software, quantum computing, e-commerce, and consumer electronics.	49. New York Skyline with 'NeurIPS' written with fireworks on the sky.
24. A large thick-skinned semiaquatic African mammal, with massive jaws and large tusks.	50. New York Skyline with 'Google Research Pizza Cafe' written with fireworks on the sky.
25. A machine resembling a human being and able to replicate certain human movements and functions automatically.	
26. A grocery store refrigerator has pint cartons of milk on the top shelf, quart cartons on the middle shelf, and gallon plastic jugs on the bottom shelf.	

Figure 1: Selected prompts for our analysis.

A.3 Attention perturbation

To assess the importance of image-to-text (I2T) attention alignment across layers, we applied a Gaussian blur along the text token dimension of the I2T attention map. We used 1D Gaussian kernel with standard deviation $\sigma = 9$ and kernel size $k = 5$ to accommodate the typical text token length. The blur was applied to the attention logits after softmax, using reflective padding to preserve the total attention mass per image token.

A.4 Segmentation evaluation metrics

To assess layer-wise segmentation performance in Sec. ??, we use three standard metrics: pixel accuracy (pACC), mean accuracy (mACC), and mean Intersection-over-Union (mIoU). Each provides a progressively more rigorous evaluation based on pixel-level true positives (TP), false positives (FP), and false negatives (FN) for the number classes N . pACC reports the fraction of correctly classified pixels over entire classes, which can be easily skewed by large classes like background. mACC averages per-class accuracy, $\frac{TP}{TP+FN}$, treating all classes equally yet still ignoring FP; mIoU is the most comprehensive, computing intersection-over-union, $\frac{TP}{TP+FP+FN}$.

$$\text{pACC} = \frac{\sum TP}{\sum TP + FN} \quad \text{mACC} = \frac{1}{N} \sum \frac{TP}{TP + FN} \quad \text{mIoU} = \frac{1}{N} \sum \frac{TP}{TP + FP + FN} \quad (1)$$

For unsupervised segmentation, we forward a noised input image with a null prompt to obtain the <pad> token attention maps. These maps are then greedily merged based on KL-divergence, following the procedure in [9]. The resulting mask proposals are evaluated via bipartite matching with ground-truth masks, while unmatched proposals are treated as false negatives.

A.5 Further details on image quality metrics.

For Pick-a-Pic, we generate 500 images for five random seeds respectively. We generated 5,000 images for MS-COCO and 1,000 images for SA-1B captions.

B Additional experiments

B.1 Ablation on timestep choice

Fig. 2 shows segmentation performance throughout different timesteps applied to the input image. Both PascalVOC and COCO-Object demonstrates the best performance on $t = 8$ of 28, where we report the segmentation performance.

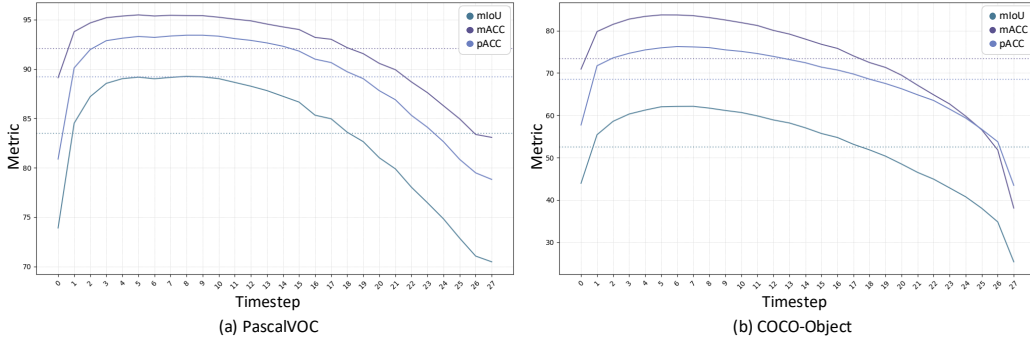


Figure 2: Segmentation performance across denoising timesteps.

B.2 Comparison on I2T and T2I attention maps

Fig. 3 presents attention maps for I2T and T2I directions corresponding to the highlighted keywords. The I2T maps exhibit more complete and contiguous object masks, whereas T2I tends to capture only partial or attenuated regions. This discrepancy likely stems from the distinct aggregation roles: I2T attention directly updates image tokens based on textual queries, while T2I updates text tokens conditioned on visual features. Given that the diffusion process ultimately operates on image tokens to synthesize outputs, it is reasonable that I2T attention aligns more strongly with semantically grounded image regions. Although this observation does not constitute a formal proof, the consistency of the qualitative patterns, which is further visualized in Appx. D.1 across prompts supports this interpretation.

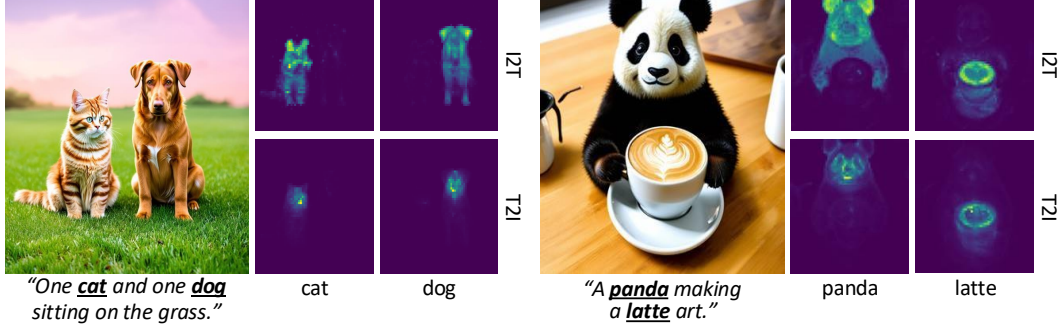


Figure 3: Comparison on I2T and T2I attention maps.

B.3 Ablation on normalization method

We conduct an ablation study to determine the optimal method for normalizing attention scores. As shown in Table 1, we evaluate the open-vocabulary semantic segmentation performance across four configurations: using scores before or after the softmax function, with and without additional min-max normalization to scale the logit between 0 to 1. The results indicate that using the raw scores directly after the softmax function yields the best performance. Consequently, we adopt this scheme for all subsequent experiments.

softmax	min-max	VOC	Object
X	X	83.2	55.1
X	✓	85.2	53.5
✓	X	89.0	61.8
✓	✓	<u>88.3</u>	<u>57.0</u>

Table 1: Ablation on attention score normalization methods.

B.4 Segmentation performance per attention layer and head

We present semantic segmentation results obtained by leveraging the attention scores from individual layers and heads in Fig.4. For a comprehensive evaluation, we measured mIoU on the PASCAL VOC dataset [3], including the background class. Our analysis reveals that the middle layers consistently exhibit superior segmentation performance compared to early or late layers.

Segmentation Performance per Head																								
Layer	Head 0	Head 1	Head 2	Head 3	Head 4	Head 5	Head 6	Head 7	Head 8	Head 9	Head 10	Head 11	Head 12	Head 13	Head 14	Head 15	Head 16	Head 17	Head 18	Head 19	Head 20	Head 21	Head 22	Head 23
Layer 0	11.0	18.5	11.7	16.0	17.8	16.4	15.2	10.2	17.8	15.9	15.7	16.3	8.8	21.2	15.2	17.8	20.5	17.6	16.6	17.9	13.4	17.7	19.6	18.2
Layer 1	16.5	12.9	15.1	19.2	16.7	0.0	0.0	21.1	15.7	15.7	13.3	8.3	19.8	0.0	13.0	5.3	18.3	17.3	17.5	10.6	13.6	14.5	16.3	9.2
Layer 2	21.5	16.9	15.7	14.2	20.8	17.0	18.1	13.5	18.3	13.5	24.7	14.8	15.3	15.1	17.8	12.8	14.8	19.5	17.1	19.1	17.7	17.1	17.1	17.5
Layer 3	0.0	17.2	19.6	18.8	20.5	0.0	0.1	23.5	21.9	12.4	30.5	19.2	17.5	32.4	0.0	17.0	19.4	0.1	27.2	16.5	15.4	18.1	0.1	23.5
Layer 4	15.3	11.8	20.5	17.9	24.1	21.2	22.3	14.7	23.1	20.4	23.2	23.0	13.3	21.8	22.9	20.1	42.4	21.1	17.3	19.0	16.5	20.5	14.2	16.7
Layer 5	30.2	34.7	32.2	18.3	26.3	17.8	26.4	36.4	22.8	23.4	23.4	21.5	24.6	32.8	29.6	18.9	19.6	14.6	25.8	21.9	29.9	26.8	22.8	22.4
Layer 6	26.8	33.2	28.6	28.8	25.5	19.2	28.1	23.1	34.2	34.6	21.4	25.0	24.9	28.2	24.0	18.1	30.1	34.9	24.6	28.3	26.7	27.6	32.2	27.8
Layer 7	38.5	51.8	26.0	40.7	42.2	33.8	35.9	28.9	21.6	24.1	31.4	33.0	17.5	34.7	31.5	45.5	30.1	24.6	29.9	16.5	31.9	36.3	29.0	32.8
Layer 8	43.5	35.9	32.0	16.2	43.1	34.4	18.2	36.4	34.3	40.2	38.3	40.5	44.5	26.1	25.3	33.6	21.4	44.1	41.9	37.2	29.3	32.5	39.0	26.3
Layer 9	23.9	37.2	25.1	31.1	29.3	28.7	32.4	22.8	28.0	28.8	26.0	24.0	22.9	26.4	27.2	26.8	32.2	32.4	37.8	32.0	31.7	25.7	37.1	24.5
Layer 10	28.8	31.6	34.3	35.8	36.3	20.9	36.4	18.3	36.6	36.7	29.8	42.4	29.9	31.3	24.1	27.6	30.9	16.0	34.4	31.8	31.5	40.9	37.1	27.4
Layer 11	32.4	30.1	25.7	30.4	33.3	26.2	19.6	25.0	26.1	25.8	21.2	31.6	37.8	31.6	28.4	34.8	33.4	28.2	31.8	22.4	33.7	25.7	29.2	27.2
Layer 12	37.6	15.8	28.4	19.8	36.5	30.0	22.9	30.7	22.1	38.7	35.3	19.5	14.4	18.5	28.7	13.1	27.4	26.5	37.0	39.6	32.2	16.0	32.5	23.6
Layer 13	34.8	33.0	25.1	34.6	31.6	26.2	35.0	29.1	31.6	33.8	35.0	31.8	35.2	39.2	29.9	30.1	25.7	31.7	31.4	35.5	32.1	28.3	27.0	34.0
Layer 14	31.5	25.3	23.3	33.2	30.6	34.1	31.9	35.6	30.1	33.3	29.5	24.9	35.7	19.1	32.9	33.9	32.2	35.6	29.2	34.1	31.7	32.5	34.0	36.7
Layer 15	23.6	23.1	27.3	30.1	25.9	26.6	29.7	29.8	24.0	23.5	30.8	29.2	27.6	29.6	31.1	29.6	28.8	26.6	28.8	27.9	25.3	25.2	28.0	24.3
Layer 16	24.1	18.4	28.6	22.9	25.1	24.5	25.7	26.2	27.4	25.3	24.2	23.0	28.9	25.4	16.4	24.5	27.7	27.3	25.2	24.4	25.2	25.9	25.1	25.8
Layer 17	30.4	24.8	24.2	20.1	26.6	25.7	26.2	20.3	17.6	25.5	24.8	26.8	25.4	26.5	25.1	21.1	24.1	21.8	27.9	26.8	26.5	25.8	26.0	25.2
Layer 18	29.6	30.1	27.2	19.8	28.7	27.4	29.2	22.4	30.8	30.6	28.5	28.2	32.5	25.2	29.9	30.6	26.7	26.1	27.6	26.4	26.3	27.4	26.8	29.1
Layer 19	28.7	28.5	29.0	26.7	32.6	14.9	29.7	24.4	20.1	29.1	27.1	28.4	23.9	14.5	29.9	23.8	30.9	16.4	30.8	30.1	24.0	25.2	25.9	30.3
Layer 20	25.5	25.4	23.7	25.1	24.9	28.4	18.7	24.2	25.4	25.2	30.3	25.7	24.7	24.9	24.3	24.4	25.5	28.1	24.1	25.5	22.4	27.0	19.0	16.5
Layer 21	27.4	23.9	25.5	27.3	26.2	20.5	27.1	26.5	21.0	24.9	25.9	23.6	24.4	29.2	26.9	27.9	25.7	25.6	23.8	26.7	26.6	19.8	23.8	19.3
Layer 22	20.2	15.9	12.9	24.6	6.7	22.4	17.9	24.5	23.3	25.8	28.1	17.4	15.5	26.8	11.7	20.7	24.3	26.1	23.8	18.1	27.5	28.3	28.2	20.2
Layer 23	25.3	24.1	21.7	18.8	28.2	18.8	21.7	21.8	21.8	19.1	17.8	20.7	23.4	28.1	24.8	21.2	23.7	26.4	24.8	28.8	20.2	24.2	23.8	26.8

Figure 4: mIoU score of each head.

B.5 Quantitative results on attention perturbation

We use CLIP-I, CLIP-T, and DINO scores to evaluate generation quality and alignment. CLIP-I measures cosine similarity between CLIP embeddings of generated and reference images, reflecting high-level perceptual fidelity. CLIP-T compares the CLIP embedding of a generated image with that of the conditioning text, assessing image–text alignment. DINO instead computes cosine similarity between self-supervised vision embeddings of generated and reference images, offering a language-free measure of semantic similarity. Higher values indicate better fidelity or alignment, with CLIP-I and DINO focusing on image–image consistency and CLIP-T on image–text consistency.

Aligned with the qualitative analysis in Sec. ??, 9th layer, which we designated as a semantic grounding expert in SD3, shows a noticeable drop on image fidelity score when perturbed.

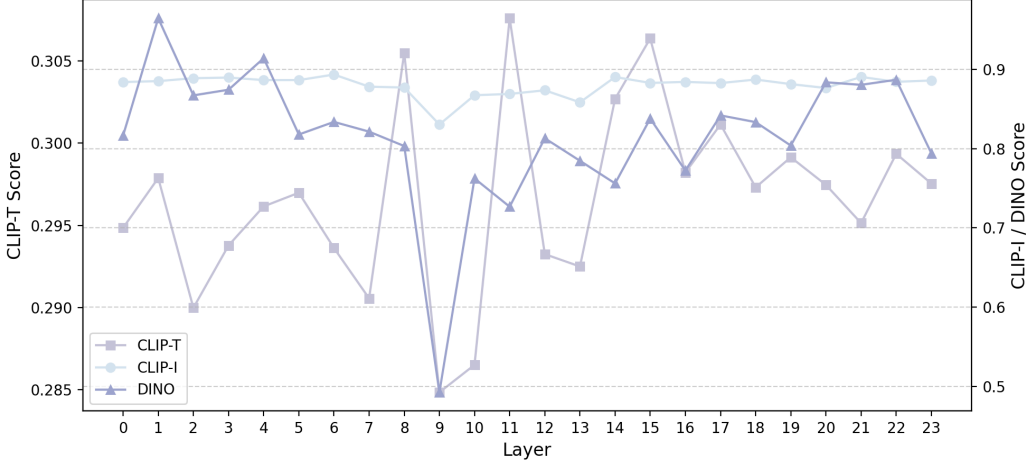


Figure 5: Image fidelity scores under layer-wise perturbations on SD3.

C Generalization to other baselines

While we mainly leverage Stable Diffusion 3 (SD3) [2] in our main analysis, we also apply our analysis to the other MM-DiT variants, Stable Diffusion 3.5 (SD3.5) [8] and Flux.1-dev [1]. We can similarly observe the correlation between value norm and segmentation performance among layers for both models. While SD3.5 appears to have layer 9, identical to SD3, to exhibit strong semantic grounding, Flux shows layer 12 and 17 to have a similar tendency. This hints that our observation and methodology are not proprietary for SD3, but can be applied to other DiT-based diffusion models with multi-modal attention, highlighting the generalizability of our insights and findings.

C.1 Stable Diffusion 3.5

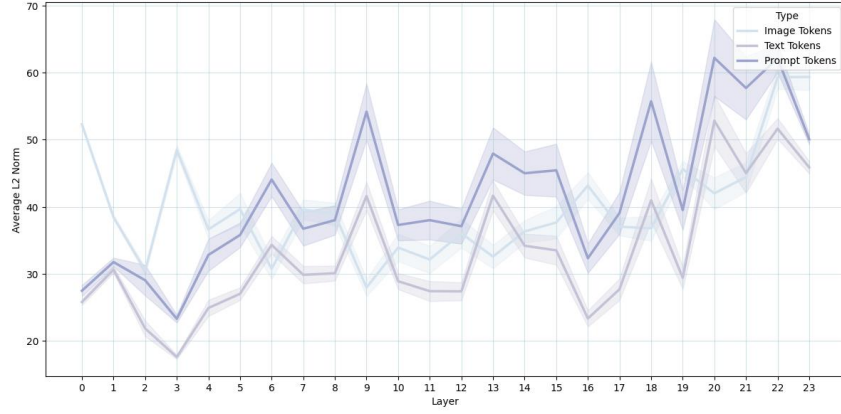


Figure 6: Average L2-Norm of values across layers on SD3.5.

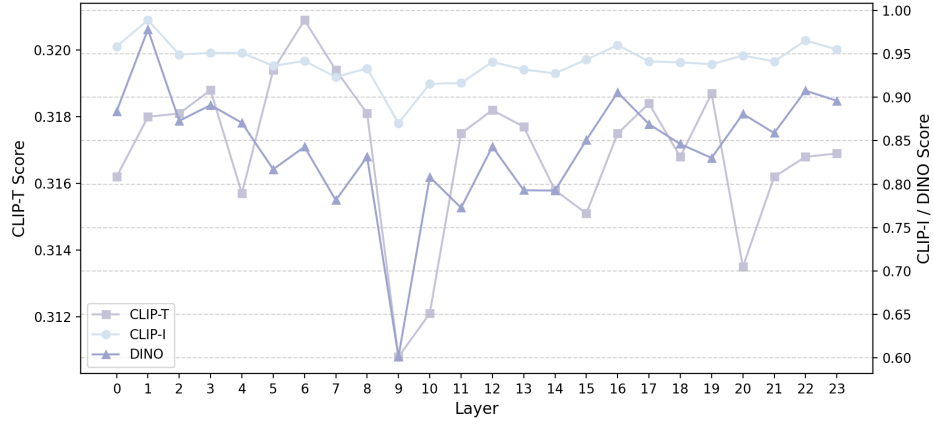


Figure 7: Image fidelity scores under layer-wise perturbations on SD3.5.

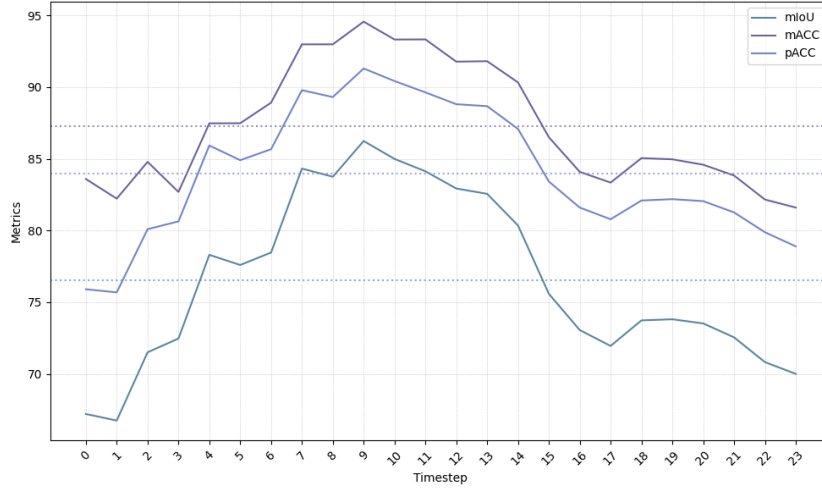


Figure 8: Segmentation performance across layers on SD3.5.

C.2 Flux.1-dev

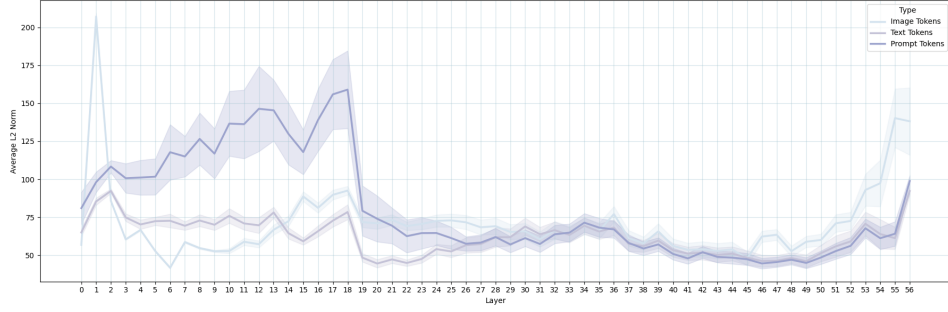


Figure 9: Average L2-Norm of values across layers on Flux.

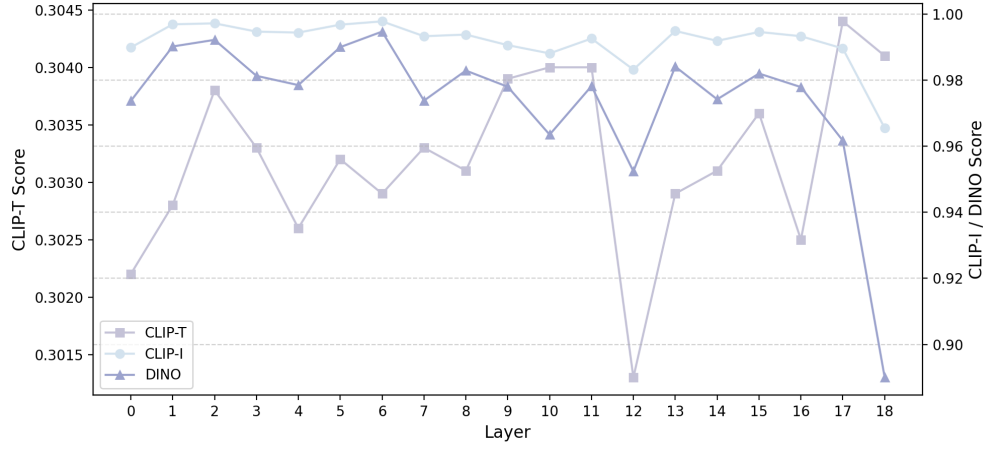


Figure 10: Image fidelity scores under layer-wise perturbations on Flux. We evaluate only for the first 19 layers, which employs MM-DiT architecture.

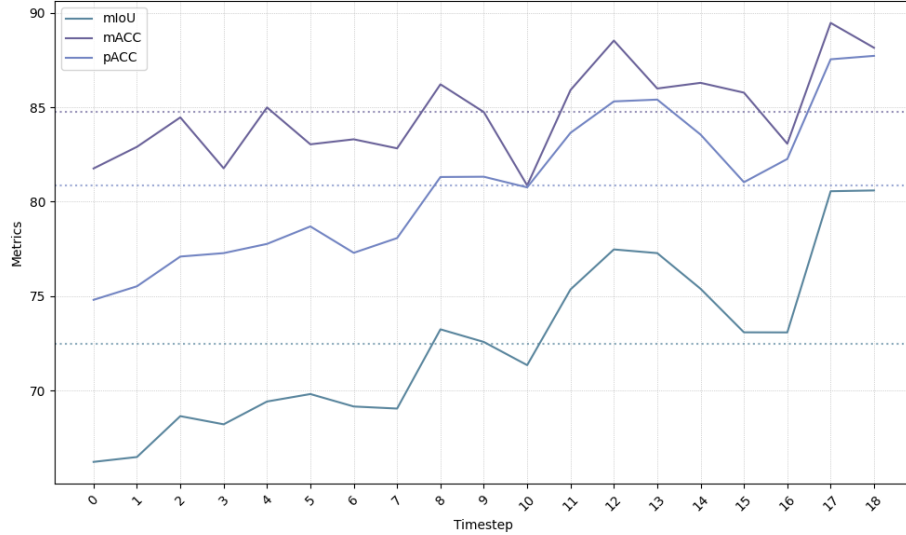


Figure 11: Segmentation performance across layers on Flux.

D Additional visualizations

D.1 Image-to-text (I2T) and text-to-image (T2I) attention map

We provide extended visualizations of image-to-text (I2T) and text-to-image (T2I) attention maps in Fig. 12. The maps are taken from layer 9, where we observe strong semantic alignment between visual and textual modalities. These results offer insight into the emergent cross-modal grounding dynamics of the multi-modal diffusion transformer (MM-DiT).

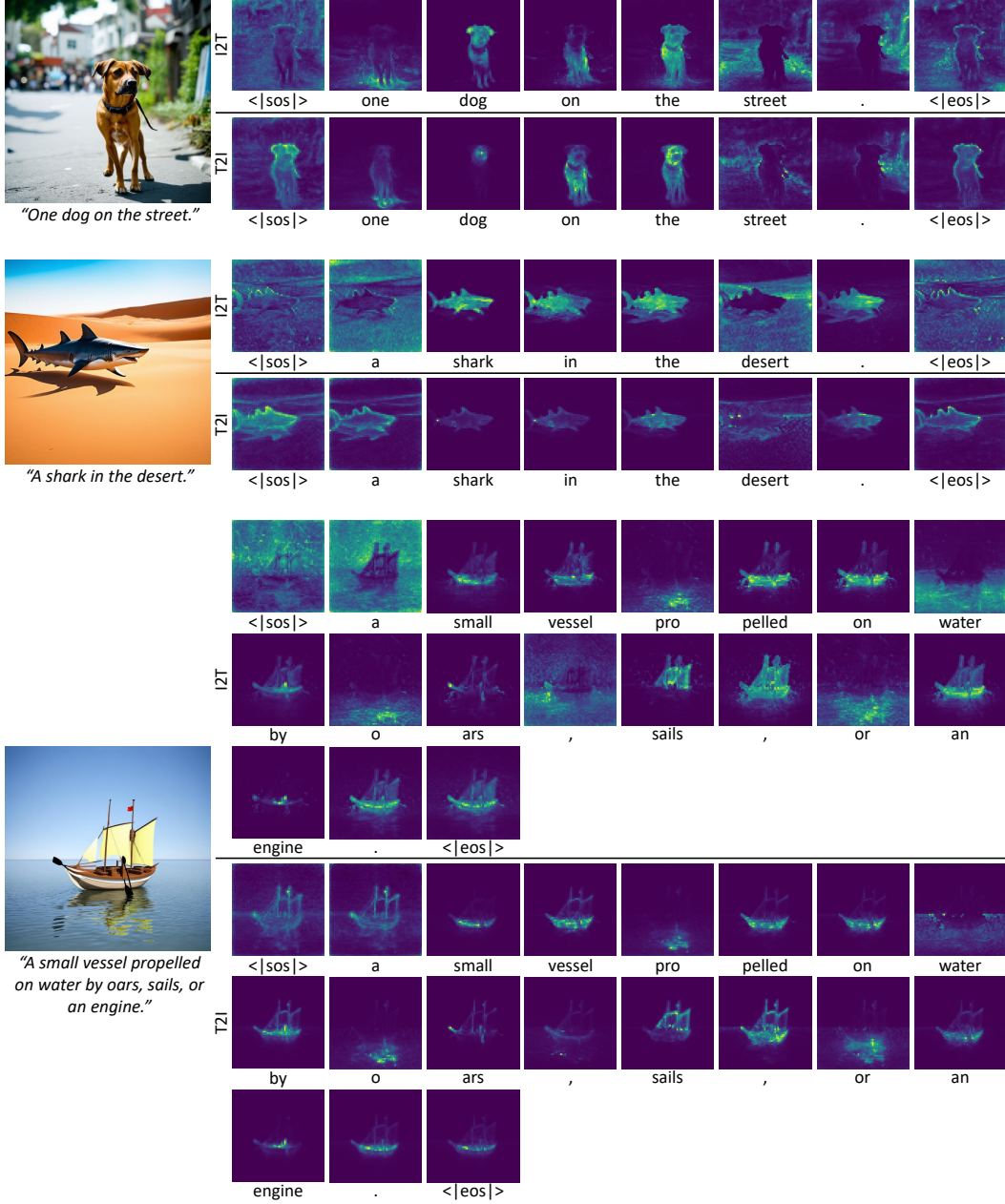


Figure 12: **Image-to-Text** attention map visualization of all prompt tokens.

D.2 Per-head attention maps

We present comprehensive visualizations of attention maps for individual heads in selected layers, as shown in Fig.13 and Fig.14. These results illustrate that each attention head focuses on distinct image regions. This effect is particularly pronounced in layer 9, where individual heads attend to different parts of the corresponding semantic region.

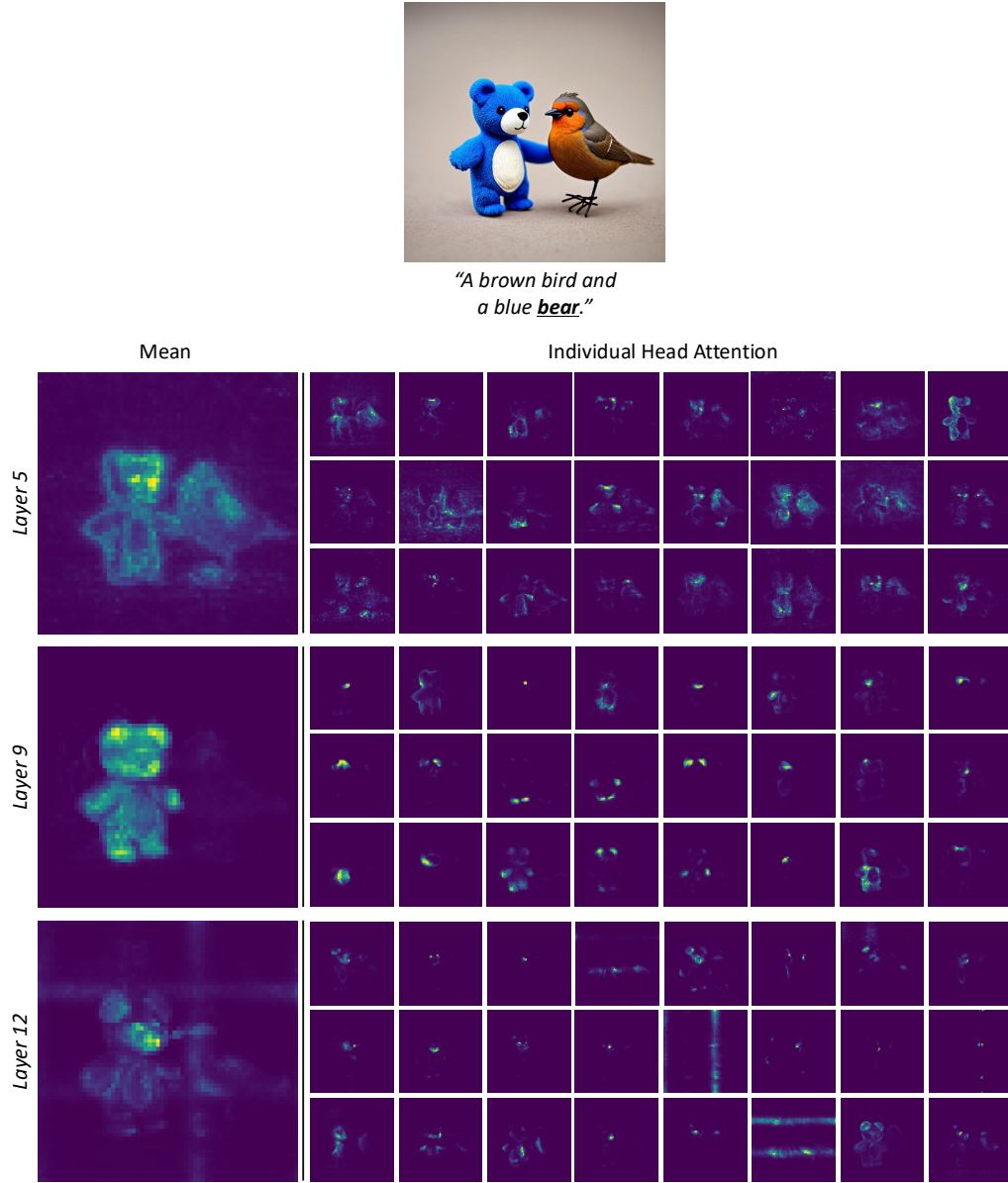


Figure 13: Visualization of attention scores for all heads.



*"Illustration of a mouse using
a **mushroom** as an umbrella."*

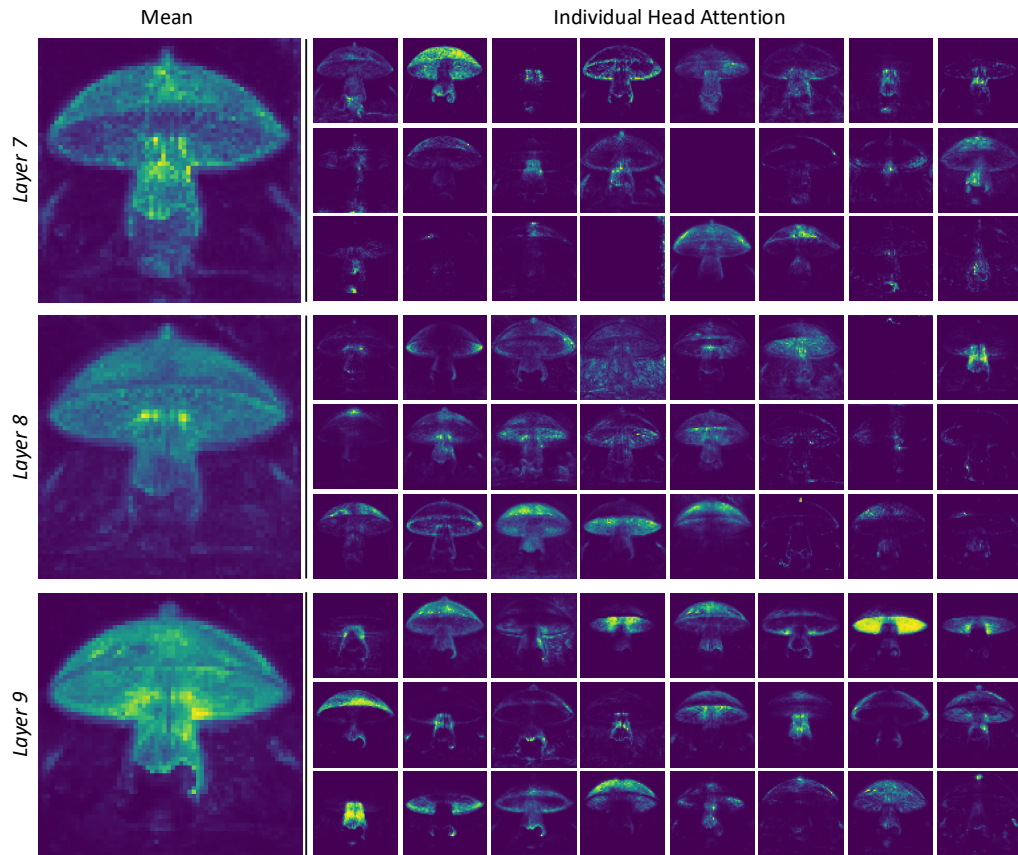


Figure 14: Visualization of attention scores for all heads.

D.3 Statistics of attention score and norm

We present the statistics of L2 norms of value-projected features across all layers and heads in Fig. 15. Notably, layer 9 exhibits consistently large norms for text tokens, suggesting that text features strongly dominate this layer. When restricting the analysis to the actual prompt tokens only, this trend becomes even more pronounced.

On the other hand, we observe large image token norms in specific heads within layers 7 and 8, which coincide with the heads that show strong segmentation performance in Fig. 4. This suggests that these heads are highly responsive to image-specific features and may play a crucial role in localizing visual semantics prior to cross-modal alignment with text.

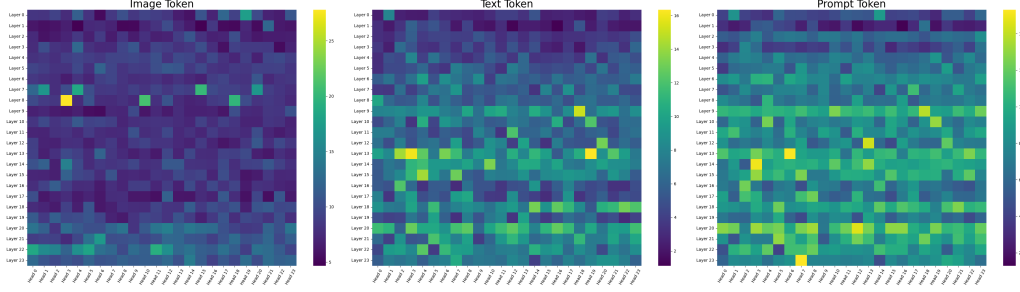


Figure 15: Average L2 norm of value tokens by layer and head.

D.4 PCA visualization of attention features

We provide PCA visualizations of the query-, key-, and value-projected image features across all layers in Fig. 16. Most layers exhibit a strong positional bias, whereas layer 9 reveals distinct semantic grouping. This suggests that image features become semantically well-grounded at layer 9, enabling more meaningful cross-modal interactions.

D.5 Emergent behavior of <pad> tokens

In Fig. 17, we visualize all 77 tokens under the unconditional generation setting described in Sec. ??, which includes <sos>, <eos>, and 75 <pad> tokens. Remarkably, we observe that individual <pad> tokens attend to distinct semantic regions, despite the absence of explicit semantic information in the text prompt.



"A bicycle on top of the boat"

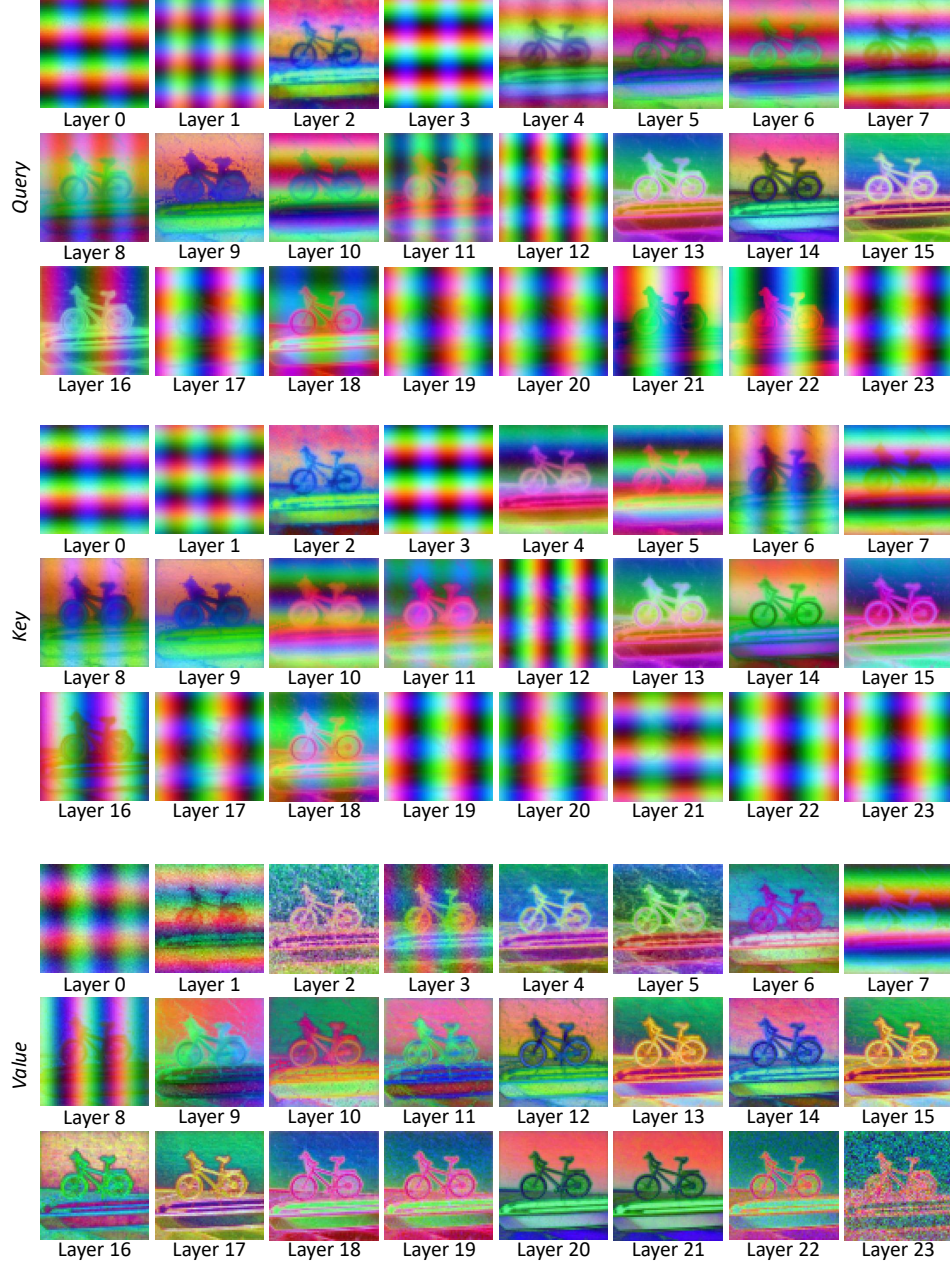


Figure 16: PCA visualization of query-, key-, and value-projected feature.



Figure 17: Emergent behavior of <pad> tokens in unconditional generation.

E Additional qualitative results for generation

We present extended results from attention perturbation experiments in Sec ?? . Perturbations applied to layers other than layer 9 result in minor degradation of image fidelity or structure while largely preserving semantic content. In contrast, perturbing layer 9 leads to the generation of semantically irrelevant images. Conversely, if we leverage this perturbed sample as a negative sample for the guidance, we observe a substantial gain on the image quality. This strongly supports our claim that layer 9 plays a critical role in cross-modal interaction, with a particular emphasis on aligning with the text modality.

E.1 Extended I2T attention perturbation results

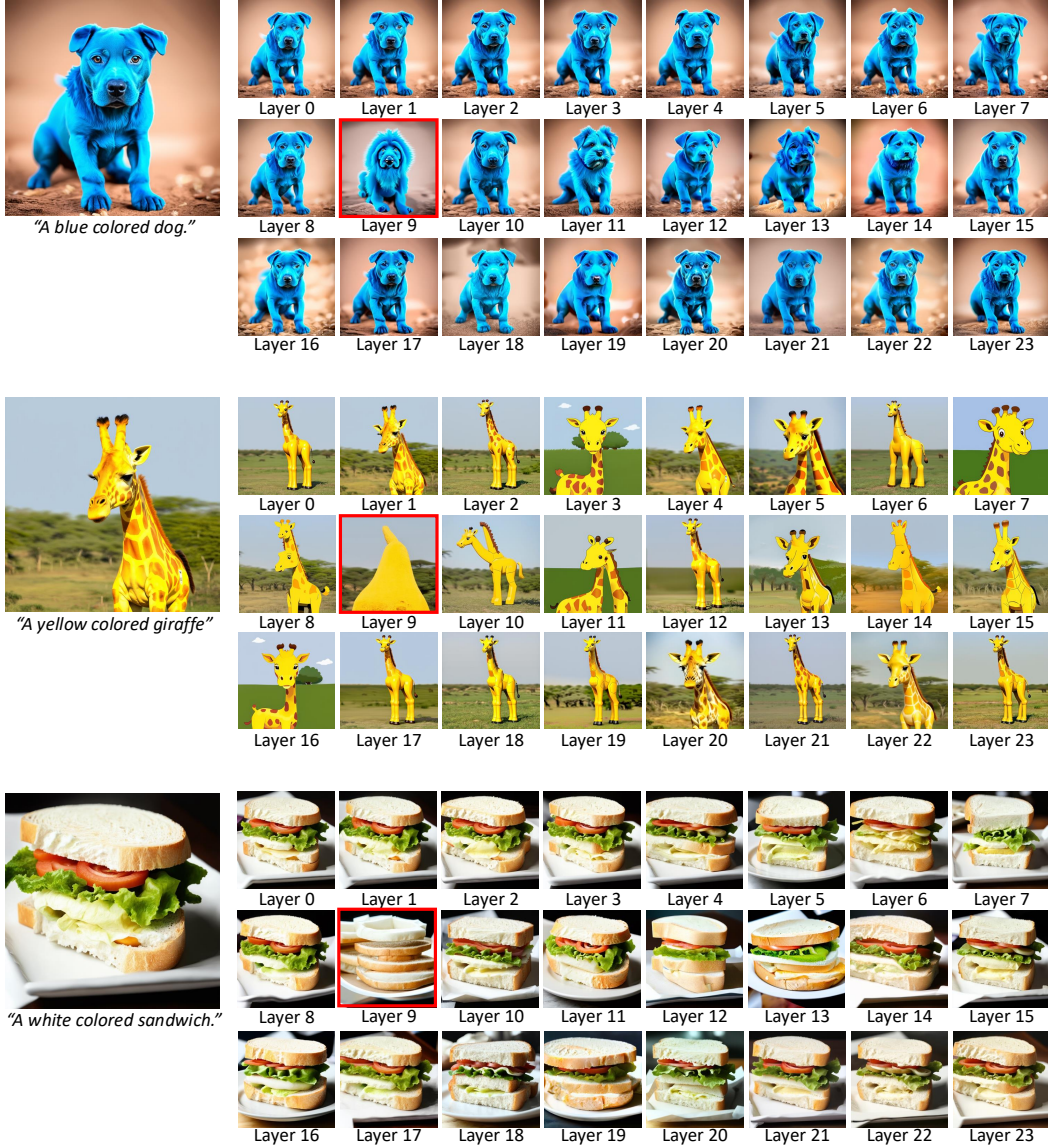


Figure 18: Effect of applying attention perturbation at different layers during image generation.

E.2 Extended I2T attention perturbation guidance results

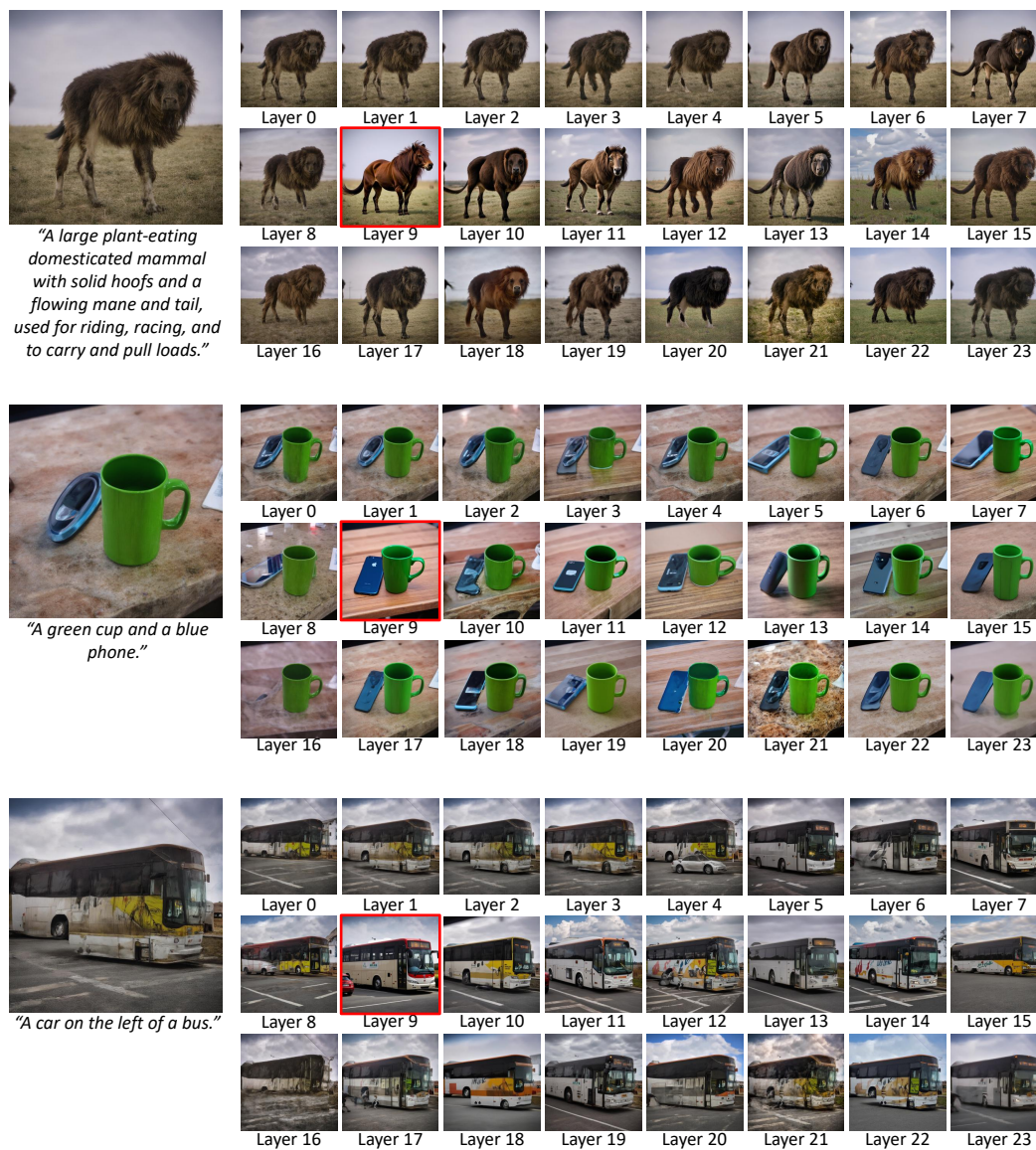


Figure 19: Effect of applying attention perturbation guidance at different layers during image generation.

E.3 Image generation results of our trained models



Figure 20: **Qualitative results of trained models.**

F Additional qualitative results for segmentation

F.1 Open-vocabulary semantic segmentation results

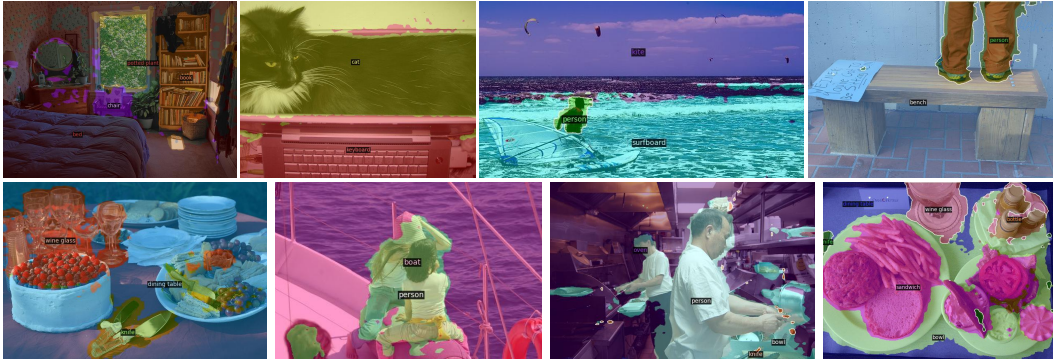
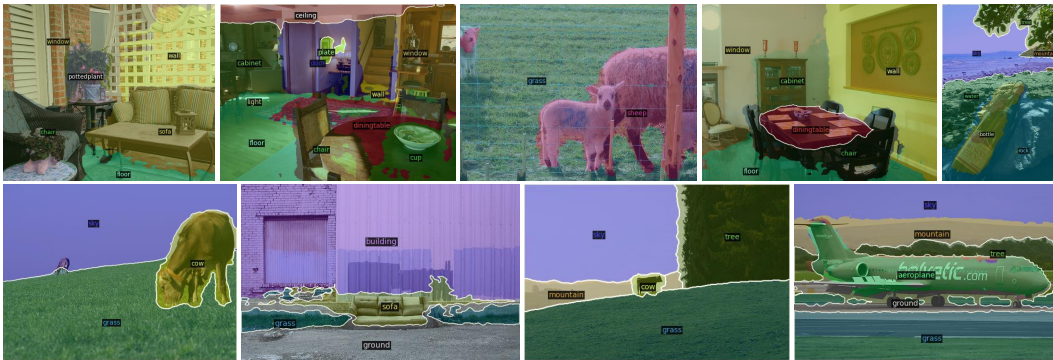
Figure 21: **Open-vocabulary semantic segmentation results of Seg4Diff on COCO-Object.**

Figure 22: Open-vocabulary semantic segmentation results of Seg4Diff on Pascal-Context59.



Figure 23: **Open-vocabulary semantic segmentation results of Seg4Diff on ADE20K.**

F.2 Unsupervised segmentation results

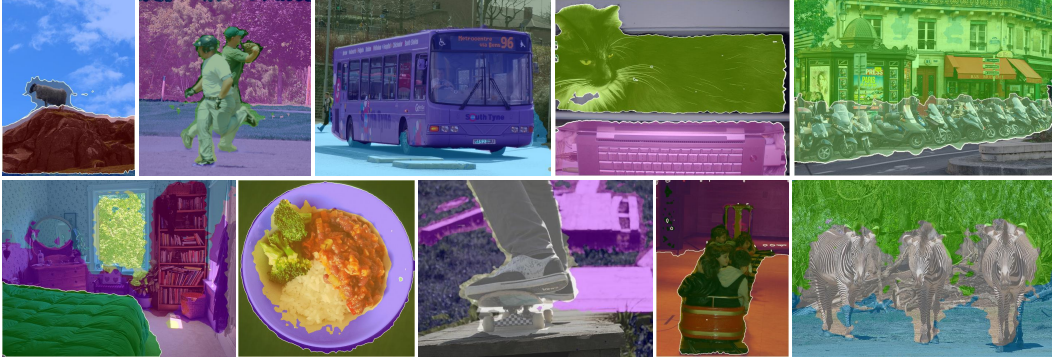


Figure 24: Unsupervised segmentation results of Seg4Diff on COCO.

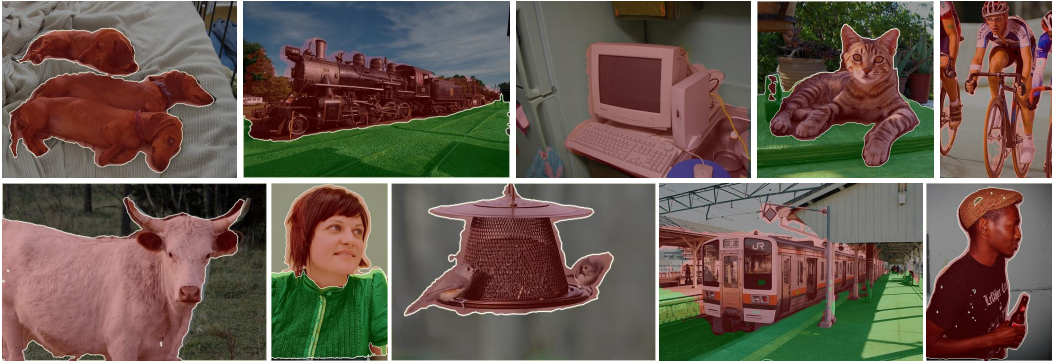


Figure 25: Unsupervised segmentation results of Seg4Diff on VOC.

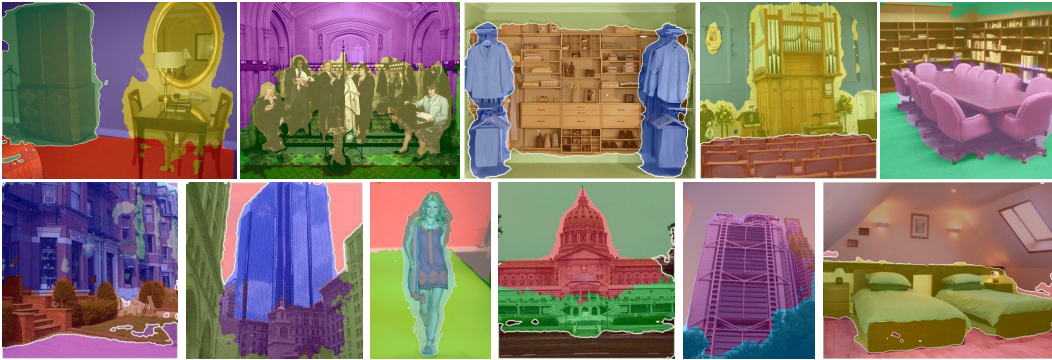


Figure 26: Unsupervised segmentation results of Seg4Diff on ADE20K.

G Limitations of our method

Our method evaluates segmentation without postprocessing or upsampling, which limits performance on extremely small objects. In addition, diffusion models may ground class names differently from the ground-truth annotations, introducing an inherent mismatch that impacts accuracy. Addressing this representation–annotation gap is an important direction for future work.

We focus on dense perception and image generation tasks, deferring reasoning-centric evaluations [10] (e.g., action recognition, spatial-relations QA) to future work, as our supervision targets segmentation semantics rather than high-level reasoning. For simplicity, the loss is applied to a single “sweet-spot” layer per backbone; broader layer/timestep exploration or auxiliary heads (e.g., optical flow, pose) could yield further gains without changing our core findings.

References

- [1] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- [3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [6] 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. *Journal of machine learning research*, 21(140):1–67, 2020.
- [7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [8] Stability AI. Stable diffusion 3.5. <https://github.com/Stability-AI/sd3.5>, 2024.
- [9] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3554–3563, June 2024.
- [10] Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.