Towards Transformer-Based Aligned Generation with Self-Coherence Guidance

Supplementary Material

6. Benchmark Details

To evaluate the capabilities of our model, we constructed a more comprehensive benchmark based on A&E [6]. Through our analysis, we identified that previous benchmarks primarily focused on coarse-grained attribute binding and lacked specific and complex scenarios. For example, prior benchmarks often evaluated prompts such as "a purple dog and a green bench." To address this limitation, we augmented the coarse-grained attribute binding tasks with specific place. For example, our prompts include cases such as "a green rabbit and a yellow bowl in the kitchen."

We argue that coarse-grained attribute binding alone is insufficient to comprehensively evaluate model performance. Therefore, we further extended the benchmark with fine-grained attribute binding and style binding tasks. Specifically, we manually created 56 fine-grained attribute binding prompts and 48 style binding prompts. Fine-grained attribute binding requires the model to control the attributes of different parts of a concept, while style prompts demand that multiple concepts within a single image exhibit distinct styles, the style prompts include categories such as "cyberpunk," "watercolor," "photorealistic," "anime," and others. The details of our benchmark is presented in Table 5.

For quantitative evaluation, we used metrics including text-to-text similarity and BLIP-VQA and additionally employed image-text similarity evaluation as discussed in section 8. Since both fine-grained attribute binding and style binding tasks involve only two concepts, we generated two questions per prompt for BLIP-VQA evaluation. For coarse-grained attribute binding tasks, as the generated image must adhere to specific locations, we generated three questions per prompt. For example, given the prompt "a blue dog and a red bench in the street," the corresponding questions are: "a blue dog?", "a red bench?", and "the street?"

This enhanced benchmark allows for a more comprehensive evaluation of model capabilities across coarse-grained, fine-grained, and style attribute bindings. Our evaluation was conducted on RTX 3090 GPUs, with each generation taking approximately 20 seconds.

7. Algorithm Details

The process of our method is detailed in Algorithm 1. Specifically, the SCG function corresponds to the approach for obtaining the new attention map described in Equation 2.

Algorithm 1 Self-Coherence Guidance.

Input: A text prompt $\mathcal{P}$, random seed s and hyper-parameter c.
Output: The latent space z_0 corresponding to images with strong consistency to the text prompt $\mathcal{P}$.

for $t = T, T-1, \ldots, 1$ **do**
  $z^*_{t-1}, A_t \leftarrow DM(z_t, t)$
  $M_t \leftarrow Cluster(A_t)/LLMplanning(A_t)$
  $h^0 = z_t$
  for $n = 1, 2, \ldots, N$ **do**
    $A^n_t \leftarrow TransBlock^n(h^{n-1})$
    $\widehat{A^n_t} \leftarrow SCG(A^n_t, M_{t+1}, c)$
    $h^n \leftarrow TransBlock^n(h^{n-1})\{A^n_t \leftarrow \widehat{A^n_t}\}$
  end for
  $z_{t-1} = h^N$
end for
Return z_0

Here, N represents the number of Transformer blocks, h represents the hidden state that each Transformer block outputs, M_t represents the masks of the concepts extracted for the next step. We iteratively replace the original attention maps with the new attention maps for each block.

8. More Quantitative Results

To further quantitatively evaluate the performance of our method, we employed image-text similarity as a metric. Following [6, 23], we utilized CLIP to separately encode images and their corresponding textual descriptions and computed their similarity scores, as shown in the Fig. 7. In this evaluation, "Full Prompts" similarity refers to the similarity between the complete prompt and the image, while "Minimum Object" similarity measures the similarity between the image and the neglected half of the text prompt.

Our method consistently outperforms previous approaches across all three tasks, with significant improvements in average similarity for both coarse-grained and fine-grained attribute binding. Notably, while the CONFORM achieves results close to ours on coarse-grained attribute binding, it fails to generalize effectively to fine-grained attribute binding and style binding, showing the poorest performance in the latter. The original PIXART-α model performs reasonably well on fine-grained attribute binding, and our approach further enhances its performance, achieving

Table 5. The details of our benchmark. BLIP-VQA refers to the number of generated questions when evaluated using the BLIP-VQA metric.

Task	Template & Example	Prompt number	BLIP-VQA
Coarse-grained	a [colorA][conceptA] and a [colorB][conceptB] *'a **black backpack** and a **pink balloon**'* a [colorA][conceptA] and a [colorB][conceptB] in the [place] *'a **blue rabbit** and a **yellow bowl** in the **kitchen**'*	54	3
Fine-grained	a [concept] with a [colorA] [partA] and a [colorB][partB] *'an apple with a **orange stem** and **blue flesh**'*	56	2
Style	a [styleA][conceptA] and a [sytleB][conceptB] *'a **anime cat** and a **photorealistic kitchen**'*	48	2

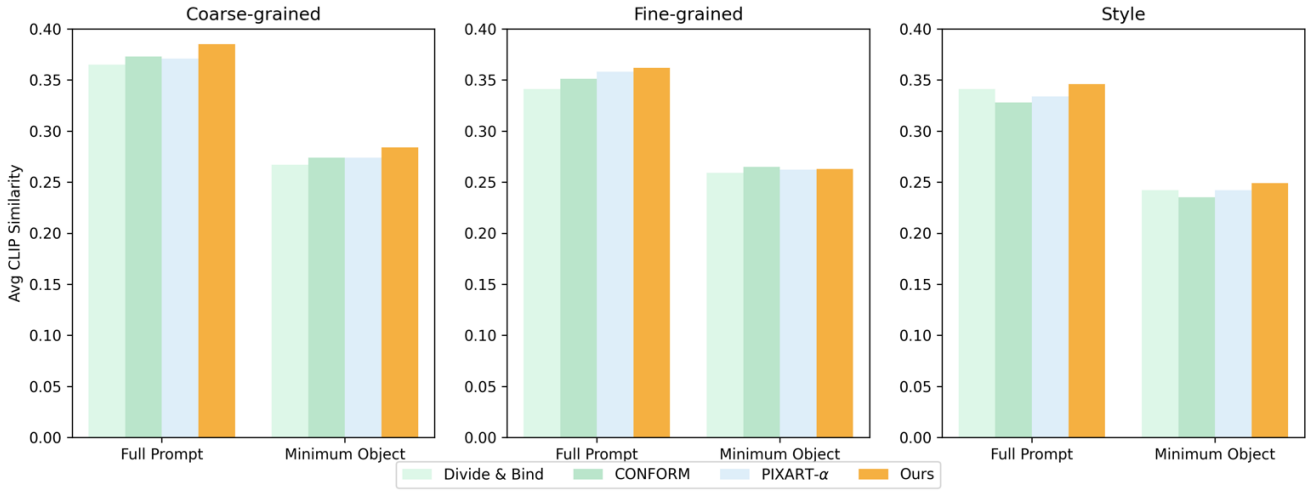

Figure 7. Comparison of average image-text similarity across different tasks. We compare our proposed method with the original PIXART-α model and two state-of-the-art aligned generation methods built on SD: D&B and CONFORM.

the best results. However, for style binding, the improvement of our method over D&B is relatively modest.

We attribute this limitation to the BLIP-caption model, which lacks specialized training for style-specific images. Consequently, it struggles to capture the fine-grained stylistic details of different concepts in images, demonstrating insensitivity to style.

In addition, we also employed the evaluation metric VQAScore. VQAScore is similar to the BLIP-VQA metric, as both assess image-text alignment by leveraging a VQA model. As shown in table 6. Our method achieves SOTA results on this additional metric as well.

To achieve a more accurate analysis, we provide a more comprehensive qualitative evaluation in the following section.

Table 6. Comparison of VQAScore. Our method still achieves state-of-the-art performance on this metric.

Method	Coarse-grained	Fine-grained	Style
D&B(SD) [19]	0.372	0.342	0.307
CONFORM(SD) [23]	0.412	0.376	0.312
PIXART-α [7]	0.387	0.437	0.320
Ours	0.476	0.488	0.366

9. Ablation Study

We conduct ablation studies to compare the performance of LLM and K-means approaches. Specifically, we evaluate both methods across three tasks: coarse-grained attribute binding, fine-grained attribute binding, and style binding. BLIP-VQA is used as the evaluation metric. The results are shown in table 7. Our experimental results demonstrate that the LLM-based approach performs better on the fine-

Figure 8. The experimental results on Flux demonstrate that our method remains effective under the MMDiT architecture..

grained attribute binding task.We attribute this to the following three reasons. **First**, the saliency of attention maps varies across different tasks, with fine-grained tasks exhibiting the least salient attention maps. **Second**, clustering algorithms process signals directly from the model output, whereas LLMs incorporate external knowledge to assist in interpreting the output. As a result, when the attention map is highly salient, clustering algorithms achieve better performance. **Finally**, LLMs provide additional benefits only when the attention map lacks saliency, making them more effective in fine-grained tasks.

Table 7. Ablation study comparing different grouping strategies.

Method	Coarse-grained	Fine-grained	Style
w/ LLM	0.647	**0.623**	0.781
w/ K-means	**0.679**	0.613	**0.799**

10. Generalizability

To verify the generalization capability of our method, we further conduct experiments on Flux [1]. Unlike PIXART-α [7], Flux [1] adopts the MMDiT architecture, which concatenates the QKV of text and image modalities before computing attention. In this setting, we still treat the dot product between the image queries and text keys as the cross-attention map on which our method operates. As shown in Fig 8, our approach remains effective under the MMDiT architecture, demonstrating strong generalization ability.

11. More Qualitative Results

To further validate the effectiveness of our approach, we provide additional qualitative analysis results for fine-grained attribute binding, style binding, and coarse-grained attribute binding tasks.These results further validate the effectiveness of our method.

Figure 9. Qualitative analysis of fine-grained attribute binding comparing our method with other SOTA approaches. Our method enables more precise control over the attributes of concepts.

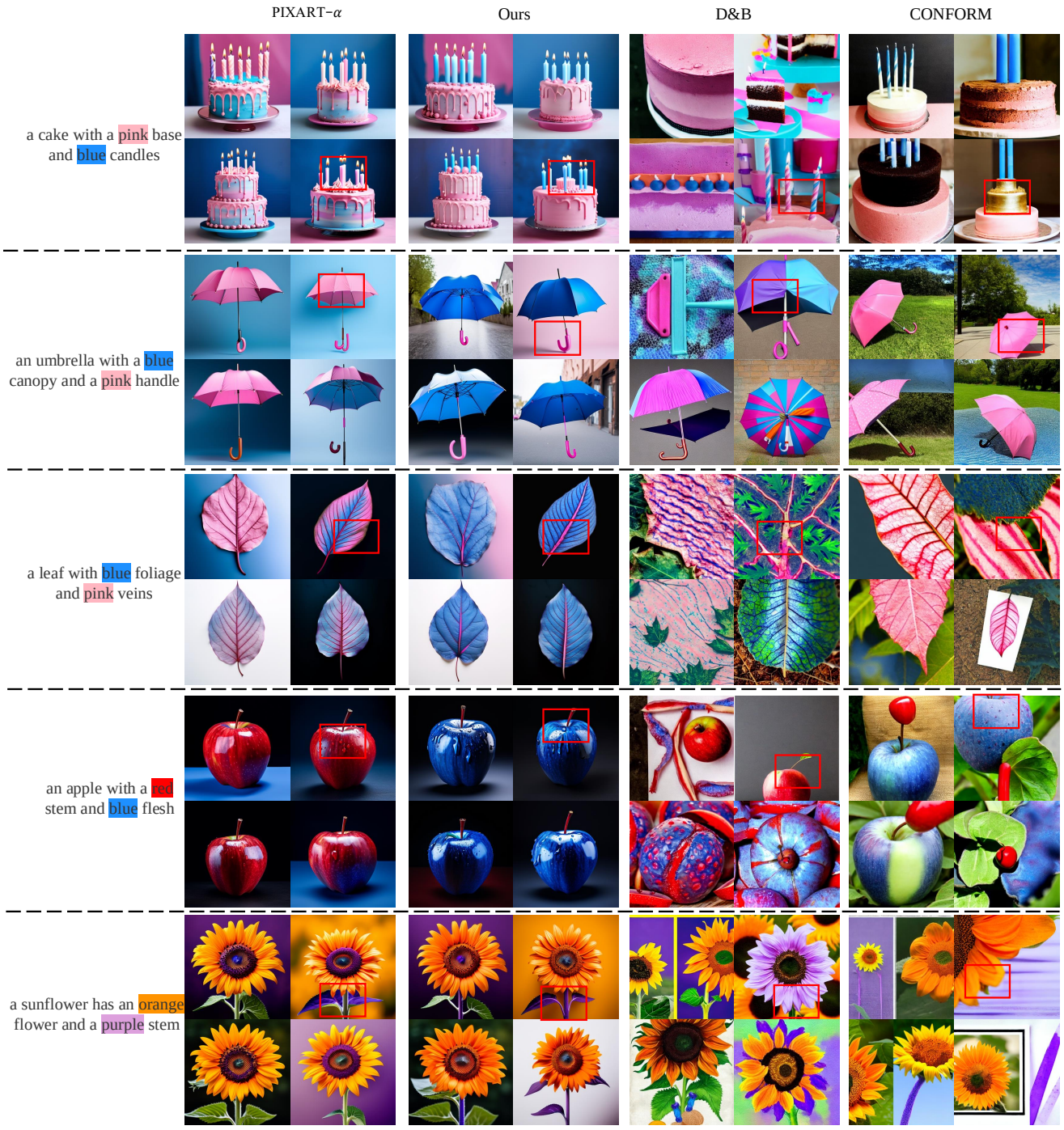

Figure 10. Qualitative analysis of fine-grained attribute binding comparing our method with other SOTA approaches. Our method enables more precise control over the attributes of concepts.

Figure 11. Qualitative analysis of style binding comparing our method with other SOTA approaches. Our method effectively binds styles to different concepts.

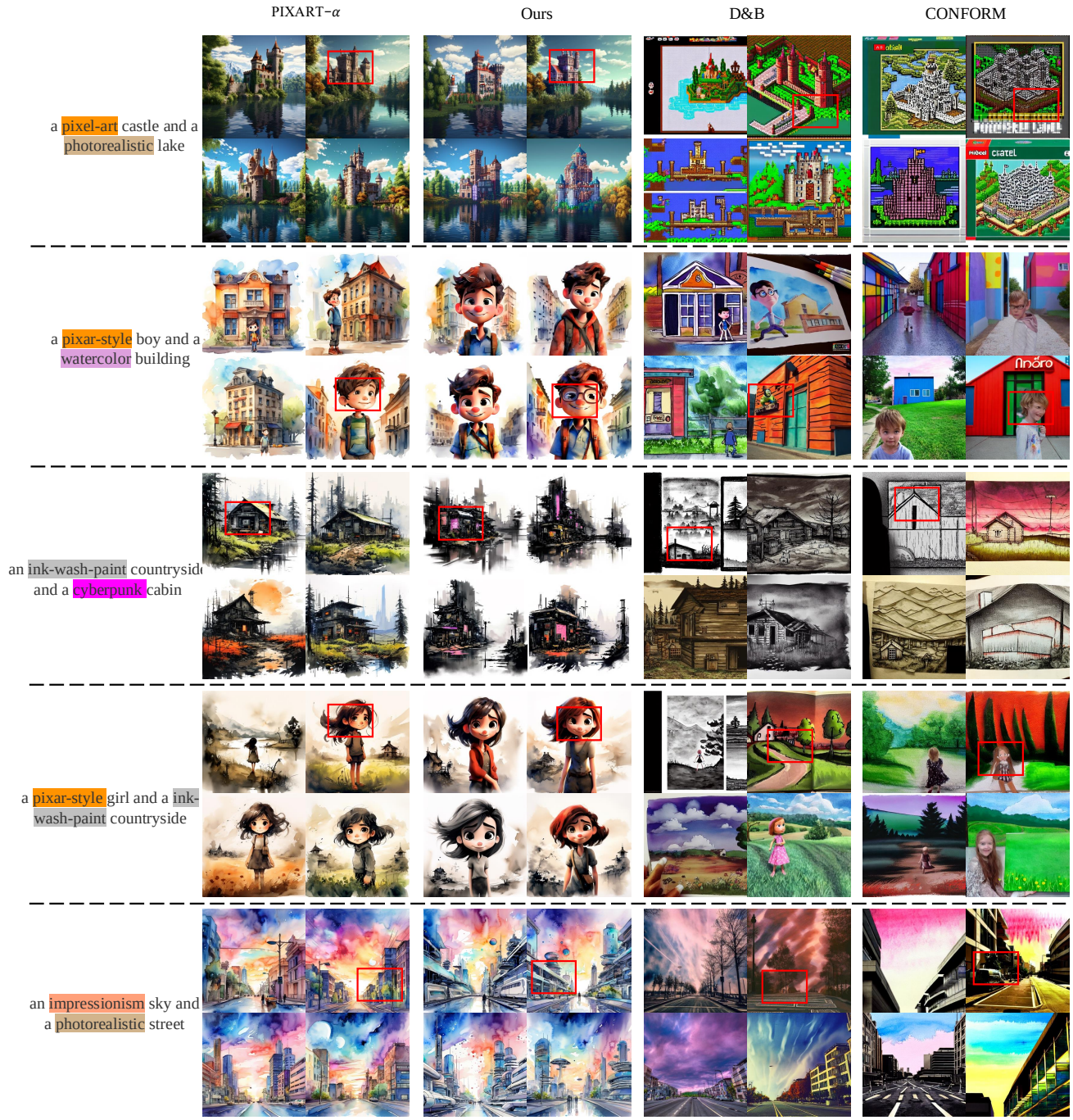

Figure 12. Qualitative analysis of style binding comparing our method with other SOTA approaches. Our method effectively binds styles to different concepts.

Figure 13. Qualitative analysis of coarse-grained attribute binding comparing our method with other SOTA approaches. Our method not only achieves attribute control but also generates higher-quality concepts.

Figure 14. Qualitative analysis of coarse-grained attribute binding comparing our method with other SOTA approaches. Our method not only achieves attribute control but also generates higher-quality concepts.

References

[1] Flux. https://github.com/black-forest-labs/flux/. 3

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 2

[4] Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. Entropy-based attention regularization frees unintended bias mitigation from lists. *arXiv preprint arXiv:2203.09192*, 2022. 7

[5] Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 2

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 3, 6, 8, 1

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2, 6, 7, 8, 3

[8] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv e-prints*, pages arXiv–2312, 2023. 2

[9] 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. *arXiv preprint arXiv:2212.05032*, 2022. 2

[10] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753, 2024. 3

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1

[12] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 3

[13] 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. 2, 5, 7

[14] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 2

[15] 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. 8

[16] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 2

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 7

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[19] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *ArXiv*, abs/2307.10864, 2023. 3, 6, 7, 8, 2

[20] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 2

[21] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 2

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[23] 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*, pages 9005–9014, 2024. 3, 6, 7, 8, 1, 2

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1

[26] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023. 5

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 2

[28] 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. 5, 7

[29] 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. 5

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[31] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[32] 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*, pages 10684–10695, 2022. 1, 2, 6

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[34] 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. 1, 2

[35] Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, et al. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *arXiv preprint arXiv:2405.15321*, 2024. 3

[36] 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. *arXiv preprint arXiv:2210.04885*, 2022. 2

[37] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 5

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[39] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 1

[41] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 2

[42] Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris Metaxas. Improving compositional text-to-image generation with large vision-language models. *arXiv preprint arXiv:2310.06311*, 2023. 2

[43] Yang Zhang, Rui Zhang, Xuecheng Nie, Haochen Li, Jikun Chen, Yifan Hao, Xin Zhang, Luoqi Liu, and Ling Li. Spdiffusion: Semantic protection diffusion for multi-concept text-to-image generation. *arXiv preprint arXiv:2409.01327*, 2024. 3

[44] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[45] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Isolated diffusion: Optimizing multi-concept text-to-image generation training-freely with isolated diffusion guidance. *arXiv preprint arXiv:2403.16954*, 2024. 2

[46] 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. *arXiv preprint arXiv:2409.19967*, 2024. 3