Selectively Extracting and Injecting Visual Attributes into Text-to-Image Models

Supplementary Material

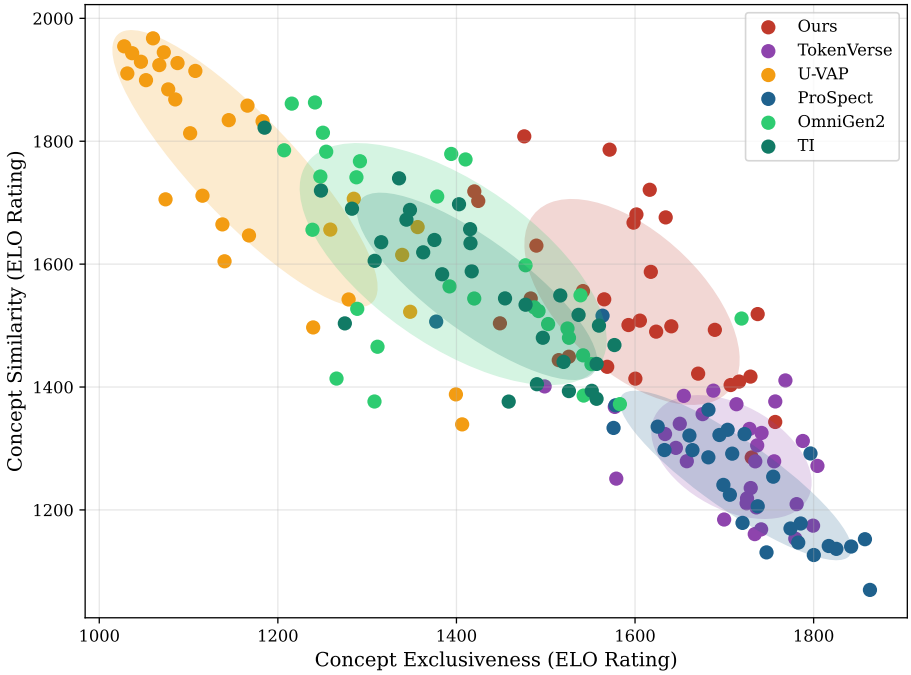


Figure 10. Elo ratings of our method and the baselines per reference image. We illustrate the distribution of each method with ellipses. The ratings are computed with a K-factor of 32 and an initial rating of 1500, following the standard Elo update rule.

8. Quantitative comparison using GPT-4o

We conduct an additional evaluation using GPT-4o to complement the quantitative results presented in the main paper. GPT-4o enables a more extensive quantitative comparison across methods, and has shown strong alignment with human judgment in recent studies [36, 43]. To perform the evaluation, we generate two images per evaluation prompt, resulting in 600 images per method. Based on the generated results, we evaluate each method's concept similarity and concept exclusiveness. We adopt a pairwise comparison framework rather than absolute scoring, as defining universal evaluation standards across diverse concept categories is inherently difficult. We compare all pairwise combinations of methods and compute Elo ratings [13] to quantify each method's relative performance.

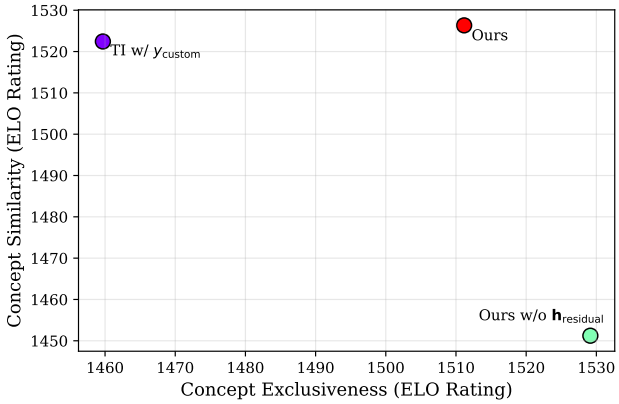


Figure 11. Elo ratings of our method and the ablation setups.

Comparison with baselines. We first prompt GPT-4o with the reference image and ask the model to describe the target concept in the image. Then, we provide two images generated by different methods and ask which image better reflects the described concept (for similarity), or which image avoids copying irrelevant attributes from the reference image (for exclusiveness). Fig. 10 visualizes the Elo ratings of all methods per reference image. The results show that U-VAP receives the lowest scores in concept exclusiveness,

Table 2. Number of iterations per reference image of our method. Check Fig. 13 to see the corresponding images.

Reference Image	# of Iterations	Reference Image	# of Iterations	Reference Image	# of Iterations
shape_vase	4,000	color_lobster	1,000	camera_wetstreet	3,000
shape_lamp	7,000	color_leather	9,000	camera_man	2,000
shape_mosque	2,000	color_tulip	4,000	camera_valencia	3,000
shape_bighead	6,000	color_bubbles	8,000	camera_lookup	10,000
shape_bridge	6,000	color_leaves	10,000	camera_boats	4,000
material_coc	1,000	pose_walk	3,000	style_binary	1,000
material_alexa	2,000	pose_pool	4,000	style_japan	2,000
material_couch	2,000	pose_kneel	3,000	style_isogame	2,000
material_glass	8,000	pose_jesus	1,000	style_poster	1,000
material_bolts	6,000	pose_swimsuit	5,000	style_childish	2,000

while TokenVerse and ProSpect achieve low concept similarity scores, which aligns with our qualitative and quantitative findings from the main paper. In contrast, our method consistently achieves high scores in both criteria, appearing in the top-right region.

Ablation studies. We also conduct the same evaluation procedure on our ablation setups. For each setup, we generate 600 images and compare them against our full method in terms of concept similarity and concept exclusiveness. Fig. 11 presents the Elo ratings of our method and the ablation setups. Unlike the ablations, our full method achieves high scores in both concept similarity and exclusiveness. This aligns with the ablation study results in the main paper and further underscore the necessity of all elements of our method to achieve accurate and controlled concept learning.

9. Image quality evaluation

To further validate that our concept learning method does not degrade the intrinsic generation quality of the base text-to-image model, we compute the Kernel Inception Distance (KID) scores. We treat the original 30 reference images in our dataset as the real distribution and evaluate the generated images across all methods. The results, computed using torch-fidelity, are presented in Table 3. Our method achieves an exceptionally low KID score, significantly outperforming TokenVerse, ProSpect, and OmniGen2, while remaining highly competitive with U-VAP and TI. These findings confirm that our approach successfully isolates and injects target concepts without compromising the overall image fidelity and naturalness.

Table 3. Quantitative comparison of image generation quality using KID scores (lower is better).

Metric	Ours	TokenVerse	U-VAP	ProSpect	OmniGen2	TI
KID ($\times 10^{-3}$)	**1.38±2.73**	5.79±3.07	**-1.50±4.12**	8.89±3.52	2.36±2.95	**1.30±2.92**

10. Extended implementation details

Ours. In addition to the implementation details in Section 5.1, we report the number of iterations per reference image in Table 2. The numbers range from 1,000 to 10,000 depending on the difficulty of the target concept. For example, extracting global styles from style-category images takes only 1,000–2,000 iterations, whereas local shapes in the shape category demand 4,000–7,000 iterations. Building on these observations, exploring automated methods to identify the appropriate number of iterations for each target concept would be a valuable direction for future research.

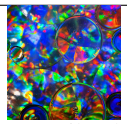
TokenVerse. As mentioned in Section 5.3, we reproduce TokenVerse on SD3 based on the details provided in the paper. The optimization process involves two stages. In the first stage, we optimize global modulation parameters for each text token for 800 iterations. During this process, we train on timesteps between 800 and 1,000 for 92% of the iterations, and on timesteps between 0 and 800 for the remaining iterations. In the second stage, we refine the modulation by optimizing a per-block offset for additional 600 iterations. In this stage, we train on timesteps between 0 and 800 for 92% of the iterations, and on timesteps between 800 and 1,000 for the rest. The learning rate for each reference image is determined by manual hyperparameter tuning among the range of 5e-6–1e-4.

11. VLM instructions for training prompt construction

We describe here the full instructions provided to the vision-language model (VLM) [20, 28, 29, 48] to obtain the custom training prompt y_{custom} and the initializer token for optimized token embedding. We provide an example of the instructions and corresponding VLM responses in Table 4.

We first provide the VLM with a reference image $\mathbf{x}_0$ and instruct the model to describe the image except for a target

Table 4. An example of instructions and corresponding VLM responses for training prompt construction.

Role	Instruction / Response
Author	 Describe the image in one detailed sentence, including the phrase "in [*]." "[*]" is a special token that already describes the colors of the bubbles in the image. Do *not* describe the colors of the bubbles in duplicate with "[*]."
VLM	The image showcases an array of iridescent bubbles in [*], set against a vibrant and swirling kaleidoscope backdrop that enhances the luminous effect.
Author	Infer the colors of the bubbles contained in "[*]" in one detailed noun phrase. Do *not* mention any elements other than the colors of the bubbles.
VLM	A spectrum of vibrant rainbow colors.
Author	Considering your previous answer, choose the best token to replace "[*]" from the following tokens. Output the token as is: "rainbow," "colorful," "colourful," "spectrum," "colors," "colours," "vibrant," "hue," "bright," "diverse."
VLM	rainbow

concept c. We also request that the model include a concept-specific phrase in the caption. These phrases are "formed in [*]" for shape, "made of [*]" for material, "in [*]" for color, "posed in [*]" for pose, "rendered in [*]" for style, and "captured in [*]" for camera shot and angle. The model then generates a caption that describes most of the non-target attributes in $\mathbf{x}_0$ while incorporating the provided phrase. The caption is descriptive enough to roughly remove the need for the embedding to learn the non-target attributes, and we use it as our custom training prompt.

Next, to select an appropriate initializer token, we ask the model to infer c that would be contained in [*] in a noun phrase. The inferred noun phrase is then used to filter candidate tokens from the entire vocabulary that have similar meanings. Specifically, we encode both the noun phrase and each token in the vocabulary into feature vectors and select the top 10 tokens whose features exhibit the highest cosine similarity to that of the noun phrase as candidate tokens. Among various publicly available encoders, we adopt a multi-QA model [38, 39] for encoding. We formulate the query as "What is the word that has the meaning of [noun phrase]?", designate each token as a potential answer, and compute the cosine similarity between the encoded representations of the query and each answer. After the filtering, we input the candidate tokens into the VLM and instruct the model to choose the best token to replace [*]. The token chosen by the model serves as a strong starting point for the optimized token embedding and is therefore used as the initializer token.

Regarding the choice of VLM, Kim et al. [22] compared several well-known image captioning models and found that GPT-4o stands out for its strong instruction-following capability. Following this finding, we adopt GPT-4o as our VLM.

12. Understanding the role of custom training prompts

To better understand the role of custom training prompts in our method, we present simple qualitative comparisons across various ablation setups in Fig. 12. On the left, the variant labeled "Ours w/o y_{custom}" uses a minimal prompt of the form "A [*]" during training. As seen in the results, the generated images exhibit a mixture of colors present in the reference image, rather than isolating the intended target color. This suggests that the distilled embedding alone is insufficient to disentangle multiple concepts within the same category, and highlights the importance of using a descriptive training prompt to guide the concept learning process.

To further examine this, we experiment with varying the level of detail in the training prompts. Specifically, we construct simplified versions of the original custom prompt (i.e., y_3) by progressively removing parts of the description generated by a vision-language model. With a moderately simplified prompt (i.e., y_2), the generated images remain largely consistent, showing that our method is robust to minor losses in prompt descriptiveness. Only when the prompt is reduced to a very minimal form (i.e., y_1), omitting more than half of the original description, do we observe occasional failures, such as the emergence of non-target colors. Nonetheless, across all cases, our method remains relatively insensitive to the level of detail in the training prompt, consistently learning the correct concept as long as a minimal degree of guidance is present.

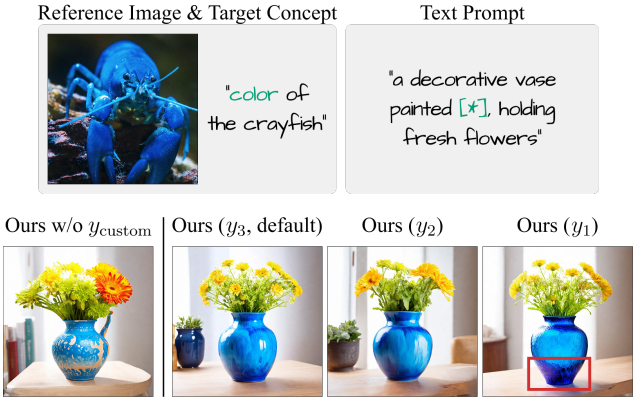


Figure 12. Qualitative comparison of generation results with varying levels of detail in the training prompt. y_3 is "a crayfish in [*] is perched on a piece of wood, its detailed exoskeleton and long antennae vividly captured against a blurred background," y_2 is "a crayfish in [*] is perched on a piece of wood, captured against a blurred background," and y_1 is "a crayfish in [*] is captured against a blurred background."

13. Full specification of reference images and evaluation prompts

We show all the reference images from our constructed dataset in Fig. 13 and the evaluation prompts in Table 5 and Table 6. The reference images are collected from Unsplash with a resolution of 1024×1024. Each exhibits unique attributes of the corresponding category. The evaluation prompts cover diverse contexts that could visually highlight the attributes of the corresponding category. The prompt design is guided by a large language model [33].

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 4

[2] Stability AI. Stable diffusion 3 medium. https://huggingface.co/stabilityai/stable-diffusion-3-medium, 2024. [Online; accessed 14-May-2025]. 6

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[4] Mário Barros and Qi Ai. Designing with words: exploring the integration of text-to-image models in industrial design. *Digital Creativity*, 35(4):378–391, 2024. 2

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. [Online; accessed 14-May-2025]. 2

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2

[7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 2

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

[9] Google DeepMind. Gemini 2.5 flash image. https://aistudio.google.com/models/gemini-2-5-flash-image, 2025. [Online; accessed 10-November-2025]. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 3

[11] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2024. 2

[12] Benj Edwards. New stable diffusion 3 release excels at ai-generated body horror. https://arstechnica.com/information-technology/2024/06/ridiculed-stable-diffusion-3-release-excels-at-ai-generated-body-horror/, 2024. [Online; accessed 14-May-2025]. 8

[13] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967. 1

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 6, 7

[15] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024. 2

[16] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025. 3, 6, 7

[17] Yuhan Guo, Hanning Shao, Can Liu, Kai Xu, and Xiaoru Yuan. Prompthis: Visualizing the process and influence of prompt editing during text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[18] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[19] Furat Jamal Hassan. Defining the design process: Methodology and creation. *Journal of Design and Textiles*, 2(1):20–35, 2023. 2

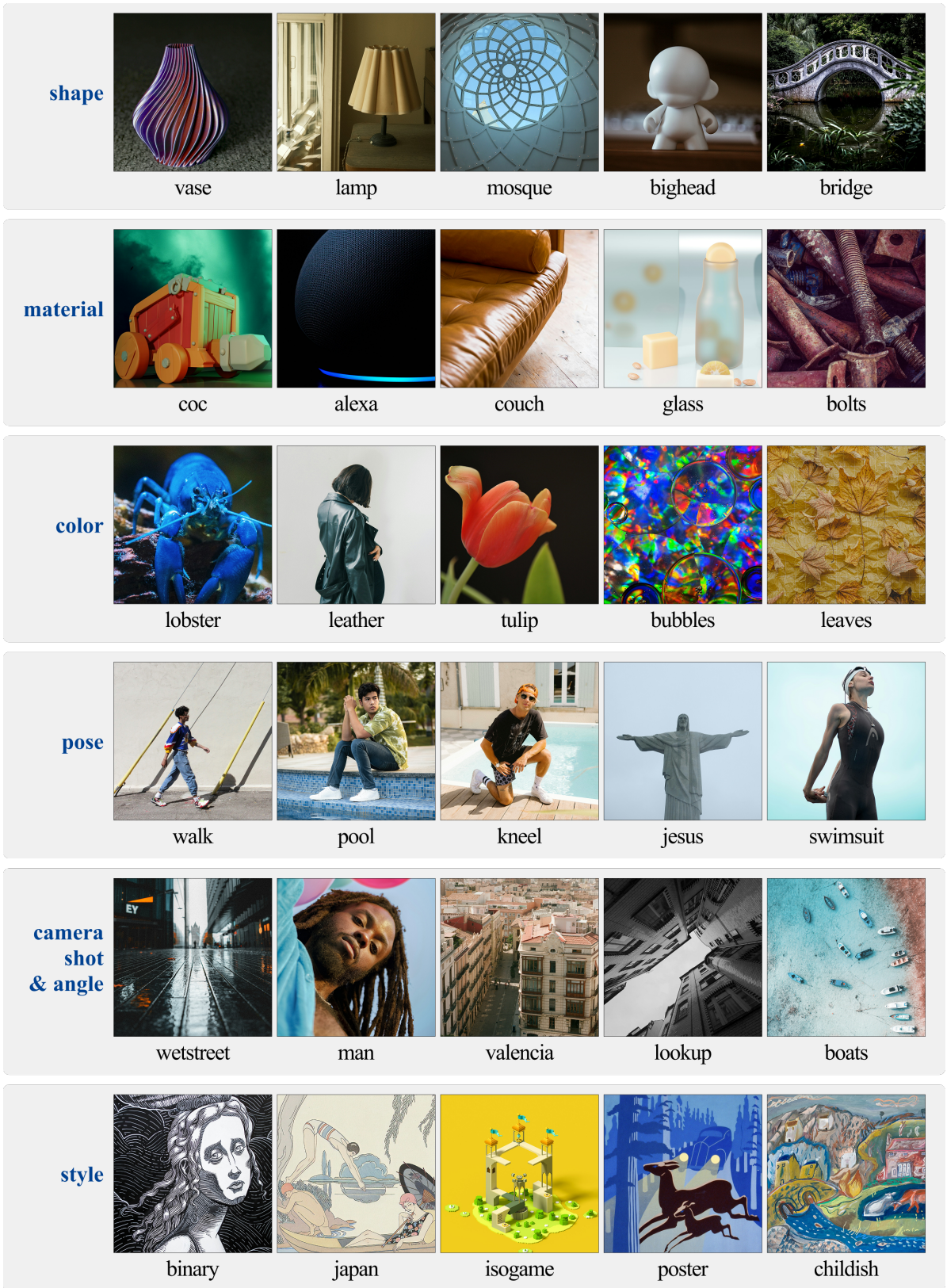


Figure 13. Reference images from our constructed dataset.

Table 5. Evaluation prompts for the shape, material, color, and pose categories.

Evaluation Prompts for Shape Category
"a stained-glass window with [*] in a cathedral, illuminated by sunlight"
"a close-up view of a gemstone cut into [*], resting on a velvet surface"
"a large cloud forming [*] in the afternoon sky"
"a pendant fashioned into [*], displayed on a black velvet stand"
"a marble sculpture carved in [*], displayed in a museum"
"a glass lamp designed in [*], placed on a table"
"an origami piece folded into [*]"
"a bold-framed mirror designed in [*], hung on a wall"
"a patio table designed in [*], placed in a sunlit garden"
"a pond shaped into [*], with stones lining its border"
Evaluation Prompts for Material Category
"a jellyfish made of [*], floating in water"
"a large daisy made of [*], set in a meadow"
"a cluster of stalactites made of [*], hanging from the ceiling of a cave"
"a row of large seashells made of [*], scattered along the shore"
"a soldier clad in armor crafted from [*], standing on the dirt ground"
"a chair made of [*], placed on a floor"
"an airplane made of [*], flying through the sky"
"a sled made of [*], set in a snowy field with mountains in the background"
"a set of wine glasses made of [*], arranged on a table"
"a three-quarter shot of a mannequin wearing a jacket made of [*]"
Evaluation Prompts for Color Category
"a jellyfish, colored [*], floating in water"
"a coffee table painted [*], placed on a rug"
"a cluster of stalactites, colored [*], hanging from the ceiling of a cave"
"a row of large seashells, colored [*], scattered along the shore"
"a soldier clad in armor painted [*], standing on the dirt ground"
"an airplane painted [*], flying through the sky"
"a product image of a wristwatch painted [*], resting on a desk"
"a set of wine glasses painted [*], arranged on a table"
"a three-quarter shot of a mannequin wearing a jacket painted [*]"
"a decorative vase painted [*], holding fresh flowers"
Evaluation Prompts for Pose Category
"a ballerina wearing a tutu, striking [*] in a studio"
"a monkey striking [*] on the grass in the afternoon sunlight"
"a surfer in a wetsuit holding [*] on a towering wave"
"a yoga practitioner holding [*] on a yoga mat"
"a dancer wearing a costume, performing [*] on a spotlight stage"
"a martial artist performing [*] inside a dojo"
"a hiker wearing a jacket, striking [*] on a mountain summit at sunrise"
"a guitarist wearing a leather jacket, striking [*] on stage"
"a child striking [*] in a playground"
"a futuristic robot striking [*] on a metallic walkway"

[20] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 3, 2

[21] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[22] Jimyeong Kim, Jungwon Park, and Wonjong Rhee. Selec-

Table 6. Evaluation prompts for the camera shot and angle, and style categories.

Evaluation Prompts for Camera Shot and Angle Category
"a city square bustling with people, captured in [*]"
"a large statue in a city square, captured in [*]"
"a row of skyscrapers captured in [*]"
"a vineyard with rows of grapevines, captured in [*]"
"a tall slide in a playground, captured in [*]"
"a skyscraper in a desert, captured in [*]"
"car lights trailing through a long-exposure effect, captured in [*]"
"two boxers in a boxing ring, captured in [*]"
"a giant waterfall captured in [*]"
"the Eiffel Tower captured in [*] at night"
Evaluation Prompts for Style Category
"a row of skyscrapers rendered in [*]"
"a hiker standing on a mountain peak, rendered in [*]"
"a marketplace with colorful stalls, rendered in [*]"
"a horse with a flowing mane, racing along a mountain trail, rendered in [*]"
"a skyscraper in a desert, rendered in [*]"
"an empty football stadium rendered in [*]"
"car lights trailing through a long-exposure effect, rendered in [*]"
"two boxers in a boxing ring, rendered in [*]"
"a helicopter landing on a roof, rendered in [*]"
"a baseball flying in a stadium, rendered in [*]"

tively informative description can reduce undesired embedding entanglements in text-to-image personalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8312–8322, 2024. 2, 3

[23] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511, 2021. 2

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1931–1941, 2023. 2

[25] Black Forest Labs. Flux.1 [dev]. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. [Online; accessed 14-May-2025]. 8

[26] Sharon Lee, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. Language-informed visual concept learning. *arXiv preprint arXiv:2312.03587*, 2023. 3

[27] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2

[28] 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. 3, 2

[29] 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. 3, 2

[30] Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. *arXiv preprint arXiv:2311.13833*, 2023. 3

[31] 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

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[33] OpenAI. Chatgpt. https://chat.openai.com/, 2022. [Online; accessed 14-May-2025]. 4

[34] OpenAI. Gpt image 1. https://platform.openai.com/docs/models/gpt-image-1, 2025. [Online; accessed 10-November-2025]. 2

[35] OpenArt. Stable diffusion prompt book. https://openart.ai/promptbook, 2022. [Online; accessed 14-May-2025]. 2

[36] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 1

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

[38] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[39] Nils Reimers and Iryna Gurevych. multi-qa-mpnet-base-cos-v1. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-cos-v1, 2021. [Online; accessed 14-May-2025]. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 22500–22510, 2023. 2

[42] 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. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2

[43] Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782, 2024. 1

[44] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 2

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. the International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 3

[46] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6): 1–13, 2023. 3

[47] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[48] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 3, 2

[49] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[50] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2, 6, 7

[51] You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, and Jintao Li. U-vap: User-specified visual appearance personalization via decoupled self augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2024. 3, 6, 7

[52] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 2

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[54] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 3, 6, 7