

Figure 9: 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.

Metric	Ours	OmniGen	U-VAP	ProSpect	TI
<b>Prompt Fidelity</b>	<u>0.782</u>	0.773	0.220	<b>0.888</b>	0.653

Table 2: Prompt fidelity of our method and the baselines.

## A QUANTITATIVE COMPARISON USING GPT-4o

We conduct an additional evaluation using GPT-4o to complement the user study 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 (Shahriar et al., 2024; Peng et al., 2024). To perform the evaluation, we generate two images per evaluation prompt across 10 prompts, resulting in 20 images per reference image and 600 images per method.

Based on the generated results, we evaluate each method along the same three criteria used in our user study: (1) concept similarity, (2) concept exclusiveness, and (3) prompt fidelity. For the first two criteria, we adopt a pairwise comparison framework rather than absolute scoring, as defining universal evaluation standards across diverse concept categories is inherently difficult. Instead, we compare all pairwise combinations of methods and compute Elo ratings (Elo, 1967) to quantify each method’s relative performance.

**Concept similarity and exclusiveness.** 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. 9 visualizes the Elo ratings of all methods per reference image. The results show that U-VAP receives the lowest scores in concept exclusiveness, while ProSpect scores lowest in concept similarity, which aligns with our qualitative findings from the main paper. Notably, our method consistently achieves high scores in both criteria, appearing in the top-right region.

**Prompt fidelity.** We prompt GPT-4o to evaluate whether each generated image correctly reflects the corresponding evaluation prompt. The results, shown in Table 2, indicate that our method achieves the highest prompt fidelity among all methods except ProSpect, which performs poorly

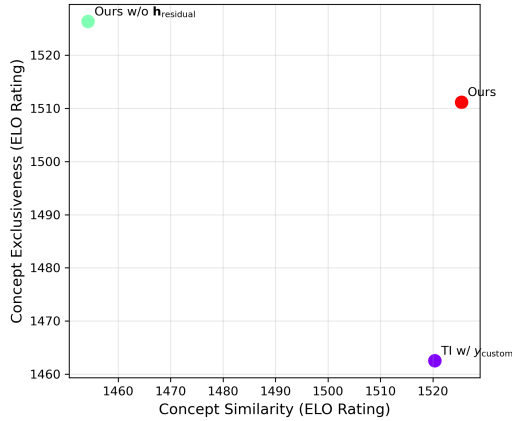


Figure 10: Elo ratings of our method and the ablation setups.

Metric	Ours	Ours w/o $h_{\text{residual}}$	TI w/ $y_{\text{custom}}$
Prompt Fidelity	<b>0.782</b>	0.723	0.605

Table 3: Prompt fidelity of our method and the ablations.

in concept similarity. Overall, these findings further demonstrate our method’s effectiveness in selectively extracting the target concept and applying it across diverse visual contexts.

**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, concept exclusiveness, and prompt fidelity. Fig. 10 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. Additionally, as shown in Table 3, our method records the highest prompt fidelity. These findings align 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.

## B EXTENDED EXPERIMENTS

### B.1 SCALING TO DIVERSE CONCEPTS

While our main experiments focus on six major categories of concepts to enable systematic evaluation, our method is not limited to these. We demonstrate its applicability to a wider range of concept types in Fig. 11.

### B.2 COMPARISON WITH GPT-4o

We also compare our method with GPT-4o. Since GPT-4o was only recently released and its API is not yet available, we generated its outputs using the ChatGPT interface. As shown in Fig. 11, we observe that GPT-4o struggles to extract implicit concepts such as camera shot and angle. In the material example below, while GPT-4o successfully reflects the target concept in its generations, it also tends to incorporate non-target attributes such as background and color. This experiment highlights the continued relevance of our concept learning task, especially in the context of increasingly capable general-purpose models being released.

### B.3 UNDERSTANDING THE ROLE OF CUSTOM TRAINING PROMPTS

To better understand the role of custom training prompts in our method, we present additional qualitative comparisons across various ablation setups in Fig. 12. On the left, the variant labeled “Ours

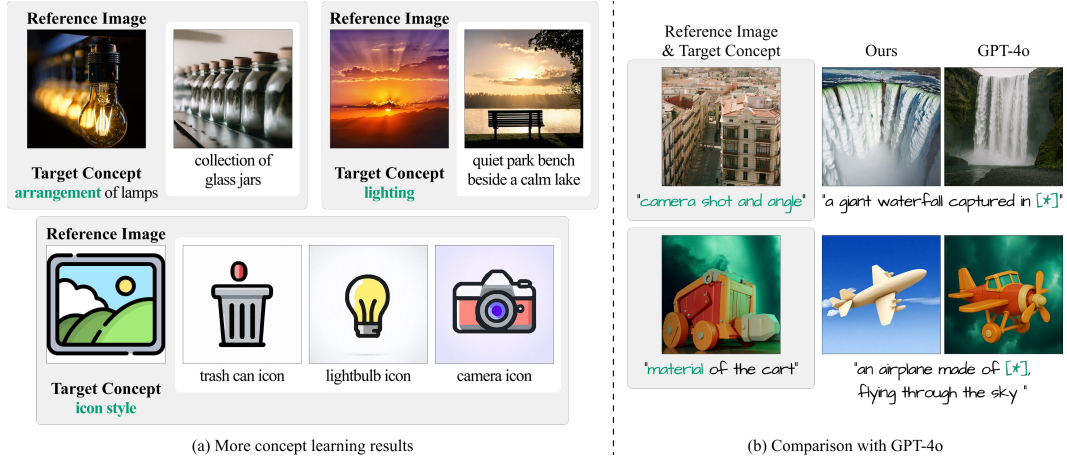


Figure 11: We provide more concept learning results and comparison with GPT-4o.

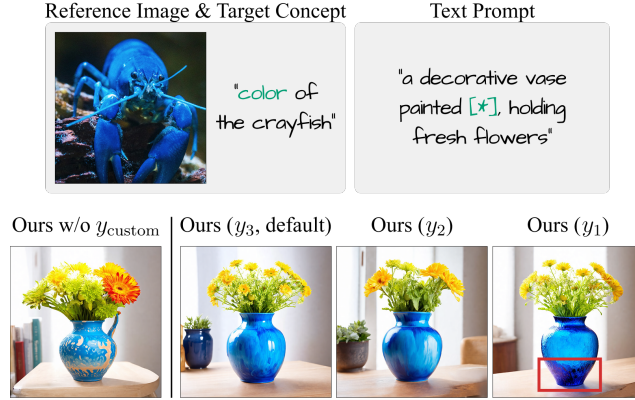


Figure 12: Qualitative comparison of generation results with varying levels of detail in the training prompt.  $y_3$  is “a crayfish in [\*] is perched on a piece of wood, its detailed exoskeleton and long antennae vividly captured against a blurred background,”  $y_2$  is “a crayfish in [\*] is perched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

w/o  $y_{\text{custom}}$ ” 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.


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

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

## C VLM INSTRUCTIONS FOR TRAINING PROMPT CONSTRUCTION

We describe here the full instructions provided to the vision-language model (VLM) (Li et al., 2022; 2023; Wang et al., 2024; Hong et al., 2024) to obtain the custom training prompt  $y_{\text{custom}}$  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  $x_0$  and instruct the model to describe the image except for a target 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  $x_0$  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 (Reimers & Gurevych, 2019; 2021) 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. (Kim et al., 2024) 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.

## D 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, each exhibiting unique attributes of the corresponding category. The evaluation prompts cover diverse contexts that could visually highlight the attributes of the corresponding category.



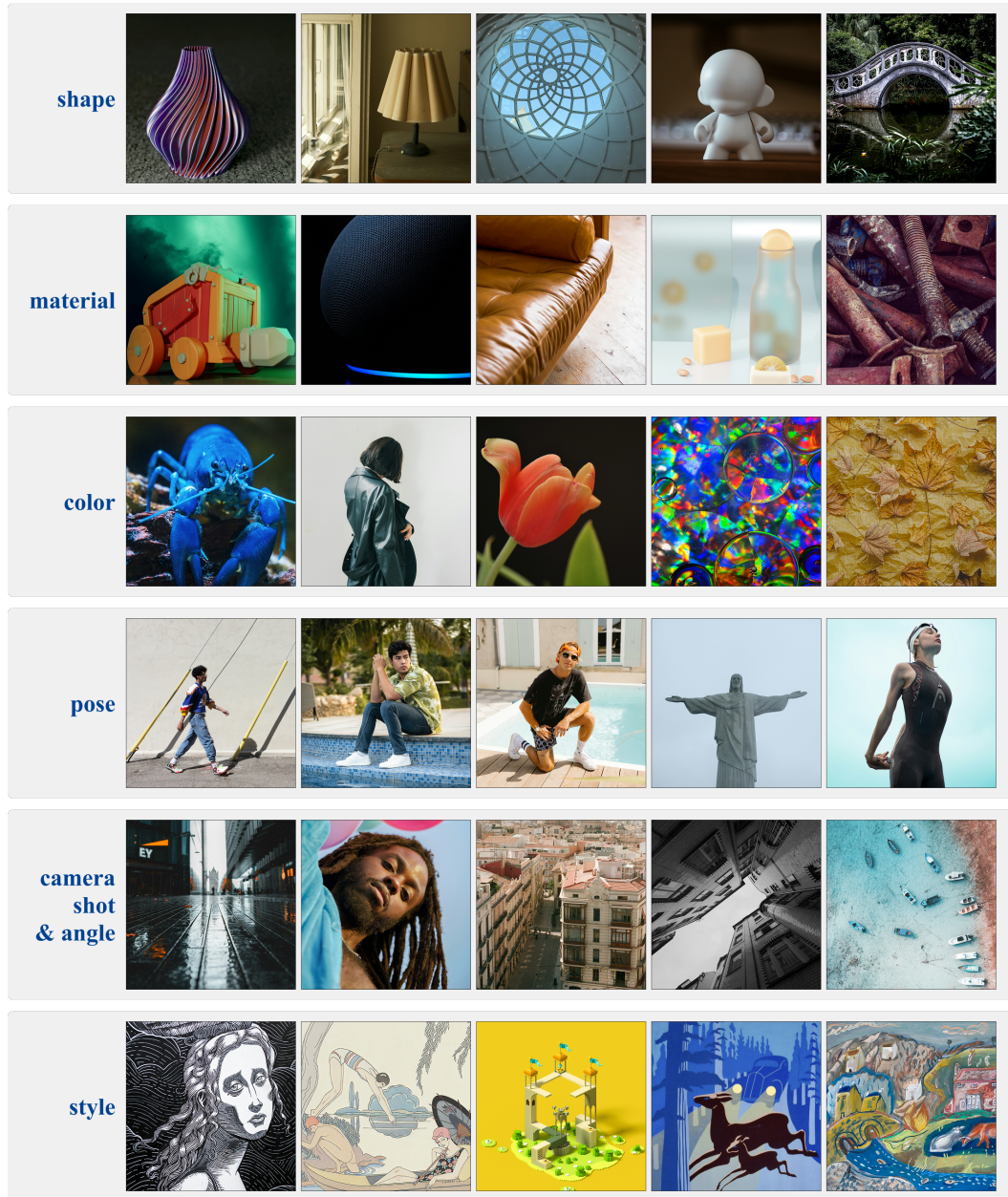


Figure 13: Reference images from our constructed dataset.

## REFERENCES

- Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967.
- Wenyi Hong, Weihang 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.
- Jimyeong Kim, Jungwon Park, and Wonjong Rhee. Selectively 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)*, pp. 8312–8322, 2024.

**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”

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

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*, pp. 12888–12900. PMLR, 2022.

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*, pp. 19730–19742. PMLR, 2023.

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.

---

**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 [\*]”

---

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

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

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].

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.

Wei Han 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.