SSR-Encoder: Encoding Selective Subject Representation for Subject-Driven Generation

Supplementary Material

In this supplementary material, we first introduce the preliminaries of Diffusion and CLIP in Section A. Following that, we provide an in-depth discussion on our Detail-Preserving Image Encoder in Section B. In subsequent sections, we introduce the methods we compared against and the user study we conducted, specifically in Section C and Section D respectively. We also present our results on human image generation in Section E. Additional results from our work on Dreambench and Multi-subject bench are showcased in Section G. We then provide further details about our training data and Multi-subject bench in Section H. In Section I and Section J, we present the outcomes generated by combining our SSR-encoder with ControlNet [24] and animatediff [5], which not only demonstrates the generalization of our SSR encoder but also illustrates its seamless applicability in the realm of controllable generation and video generation for maintaining character consistency with reference images. Lastly, we analyze the broader impact brought by our method and the limitation of our method in Section K and Section L.

A. Preliminaries

A.1. Preliminary for Diffusion Models

Diffusion Model (DM) [7, 21] belongs to the category of generative models that denoise from a Gaussian prior $\mathbf{x_T}$ to target data distribution $\mathbf{x_0}$ by means of an iterative denoising procedure. The common loss used in DM is:

$$L_{simple}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x_0},t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}\left(\mathbf{x_t}, t\right)\|_2^2\right], \quad (1)$$

where $\mathbf{x_t}$ is an noisy image constructed by adding noise $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ to the natural image $\mathbf{x_0}$ and the network $\boldsymbol{\epsilon_\theta}(\cdot)$ is trained to predict the added noise. At inference time, data samples can be generated from Gaussian noise $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ using the predicted noise $\boldsymbol{\epsilon_\theta}(\mathbf{x_t}, t)$ at each timestep t with samplers like DDPM [7] or DDIM [20].

Latent Diffusion Model (LDM) [17] is proposed to model image representations in autoencoder's latent space. LDM significantly speeds up the sampling process and facilitates text-to-image generation by incorporating additional text conditions. The LDM loss is:

$$L_{LDM}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x_0},t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}\left(\mathbf{x_t}, t, \boldsymbol{\tau_\theta}(\mathbf{c})\right)\|_2^2\right], \quad (2)$$

where $\mathbf{x_0}$ represents image latents and $\boldsymbol{\tau_\theta}(\cdot)$ refers to the BERT text encoder [3] used to encodes text description $\mathbf{c_t}$.

Stable Diffusion (SD) is a widely adopted text-to-image diffusion model based on LDM. Compared to LDM, SD is trained on a large LAION [19] dataset and replaces BERT with the pre-trained CLIP [16] text encoder.

A.2. Preliminary for CLIP

CLIP [16] consists of two integral components: an image encoder represented as $F(x)$, and a text encoder, represented as $G(t)$. The image encoder, $F(x)$, transforms an image x with dimensions $\mathbb{R}^{3\times H\times W}$ (height H and width W) into a d-dimensional image feature f_x with dimensions $\mathbb{R}^{N\times d}$, where N is the number of divided patches. On the other hand, the text encoder, $G(t)$, creates a d-dimensional text representation gt with dimensions $\mathbb{R}^{M\times d}$ from natural language text t, where M is the number of text prompts. Both encoders are concurrently trained using a contrastive loss function that enhances the cosine similarity of matched pairs while reducing that of unmatched pairs. After training, CLIP can be applied directly for zero-shot image recognition without the need for fine-tuning the entire model.

B. Designing Choice of Image Encoder

In this section, we conduct a preliminary reconstruction experiment to demonstrate that vanilla image features fail to capture fine-grained representations of the target subject and verify the effectiveness of our method. We first introduce our experimental setup and evaluation metrics in Sec. B.1. Subsequently, we explain the implementation details of each setting in Sec. B.2. Finally, we conduct qualitative and quantitative experiments in Sec. B.3 to prove the superiority of our proposed methods compared to previous works.

B.1. Experimental Setup

In our image reconstruction experiment, we investigate four types of image features. The details are as shown in Fig. 1:

- **Setting A: CLIP Image Features.** In this setting, we employ the vanilla CLIP image encoder to encode the input image and utilize the features from the final layer as the primary representation for subsequent reconstruction.
- **Setting B: DINOv2 Image Features.** Analogous to setting A, we replace the CLIP image encoder with the DINOv2 encoder to extract the features.
- **Setting C: Fine-tuned CLIP Image Features.** With the goal of recovering more fine-grained details while preserving text-image alignment, we fine-tune the last layer

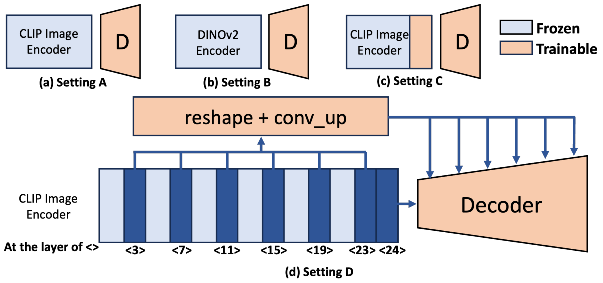


Figure 1. Details for each setting.

parameters of the CLIP image encoder using a CLIP regularization loss.

- **Setting D: Multi-scale CLIP Image Features.** Instead of fine-tuning, we resort to using features from different scales of the CLIP backbone as the image representations.

To verify the effectiveness of our methods, we employ the following metrics: **Perceptual Similarity (PS)** [25] and **Peak Signal-to-Noise Ratio (PSNR)** to assess the quality of reconstruction, **CLIP-T** [6] and **Zero-Shot ImageNet Accuracy (ZS)** [2] to access the preservation of text-image alignment in image encoder variants.

As for data used in our preliminary experiments, we utilize a subset of LAION-5B [19]. This dataset comprises approximately 150,000 text-image pairs for training and a further 10,000 text-image pairs designated for testing.

B.2. Implementation Details

We use OpenCLIP ViT-L/14 [8] and DINOv2 ViT-L/14 [15] as the image encoders and all images are resized to 224×224 for training. The model underwent 100,000 training iterations on 4 V100 GPUs, using a batch size of 32 per GPU. We adopt the Adam optimizer [9] with a learning rate of 3e-4 and implement the one-cycle learning scheduler. To better preserve the pre-trained weights, we set the learning rate of the image encoder as 1/10 of the other parameters if fine-tuning is required. We adopt the same architecture of the VAE decoder in LDM [17] with an extra upsampling block and employ nearest interpolation to obtain the final reconstruction results. We adopt L_2 reconstruction loss in all our settings and additionally employ L_{clip} when fine-tuning the CLIP encoder.

B.3. Experiment Results

Qualitative results. To demonstrate the effectiveness of our method, we present reconstruction results in Fig. 2. It is observed that vanilla CLIP image features and DINOv2 features only result in rather blurry outcomes. By contrast, both fine-tuned CLIP image features and multi-scale CLIP image features manage to retain more details. Specifically, multi-scale CLIP image features is able to generate sharp edges without obvious degradations. Consequently, we infer that multi-scale features are more competent at preserving the fine-grained details we require.

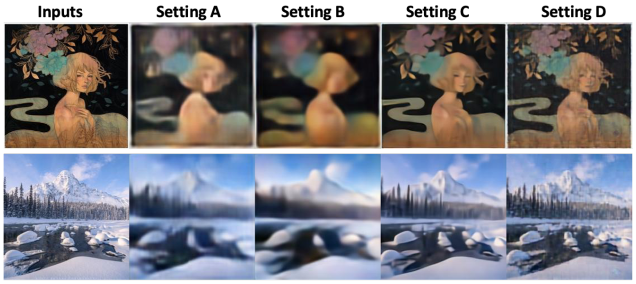


Figure 2. Comparisons of different settings.

Quantitative results. The quantitative results are shown in Table 1. In terms of reconstruction quality, it's noteworthy that both the fine-tuned CLIP image features and multi-scale CLIP image features are adept at producing superior outcomes, exhibiting lower perceptual similarity scores and higher PSNR. This indicates that these features are more representative than either vanilla CLIP image features or DINOv2 features. However, despite the assistance from CLIP regularization loss, fine-tuned CLIP image features still suffer significant degradation in text-image alignment, which fails to meet our requirements. Consequently, we opt for multi-scale features as our primary method for extracting subject representation.

Table 1. Comparisons of different settings.

Settings	PS ↓	PSNR ↑	CLIP-T ↑	ZS ↑
A	0.0036	28.63	**0.1816**	**75.3%**
B	0.0013	28.56	–	–
C	**0.0004**	**29.73**	0.1394	68.4%
D	0.0006	29.49	**0.1816**	**75.3%**

C. Details of Comparison Experiments

C.1. Details of Compared Methods

1. **Finetune-based Methods:**
   - **Textual Inversion** [4]: A method to generate specific subjects by describing them using new "words" in the embedding space of pre-trained text-to-image models.
   - **Dreambooth** [18]: A method of personalized image generation by fine-tuning the parameters in diffusion U-Net structure.
   - **Break-A-Scene** [1]: Aims to extract a distinct text token for each subject in a single image, enabling fine-grained control over the generated scenes.
2. **Finetune-free Methods:**
   - **Reference only** [14]: Guide the diffusion directly using images as references without training through simple feature injection.
   - **ELITE** [22]: An encoder-based approach encodes the visual concept into the textual embeddings for subject-driven image generation.

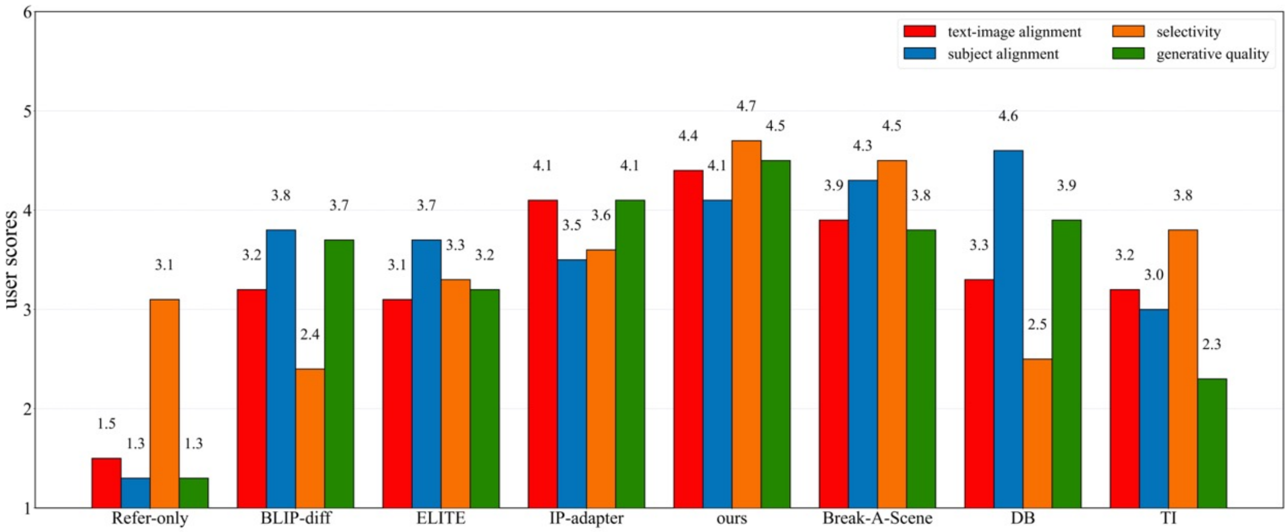


Figure 3. User study comparisons of different methods.

- **IP-adapter** [23]: Focuses on injecting image information without fine-tuning the base model.
- **BLIPDiffusion** [11]: Combines BLIP's language-image pretraining with diffusion models.

These methods were chosen for their relevance and advancements in the field, providing a robust frame of reference for evaluating the performance and innovations of our SSR-Encoder.

C.2. Details of Implementation

In order to achieve a fair comparison, all the methods are implemented using the official open-source code based on SD v1-5 and the official recommended parameters. For the Multi-subject bench, all the methods use a single image as input and utilize different subjects to guide the generation. We provide 6 different text prompts for each subject on each image and generate 6 images for each text prompt. For Dreambench, we follow [11, 18] and generate 6 images for each text prompt provided by DreamBench.

D. User Study

We conducted a user study to compare our method with DB, TI, Break-A-Scene, ELITE, and IP-adapter perceptually. For each evaluation, each user will see one input image with multiple concepts, two different prompts for different concepts, and 5 images generated by each prompt and each method. 60 evaluators were asked to rank each generated image from 1 (worst) to 5 (best) concerning its selectivity, text-image alignment, subject alignment, and generative quality. The results are shown in Table. 3 indicate that our method outperforms the comparison methods in generative quality and better balances subject consistency and text-image alignment.

E. Human Image Generation

Despite the SSR-Encoder not being trained in domain-specific settings (such as human faces), it is already capable of capturing the intricate details of the subjects. For instance, similar to the method outlined in [13], we utilize face images from the OpenImages dataset [10] as reference images for generating human images. Fig. 5 showcases samples of the face images we generated. To better illustrate our results, we also employ images of two celebrities as references.

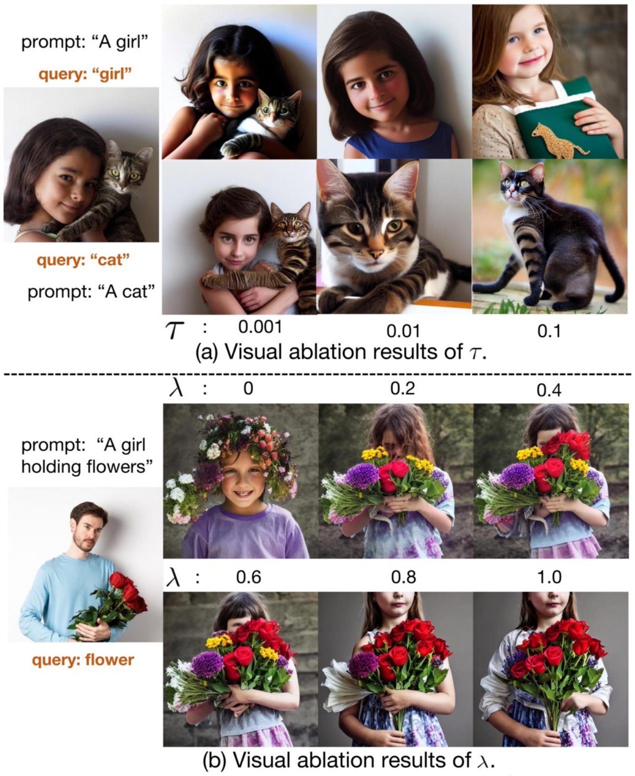
F. Ablations of τ and λ

As shown in Fig. 4 (a), under the same training settings, when τ was 0.01, the model managed to balance both identity consistency and selectivity. The effects of different λ values on the images under ablation and fixed seed conditions are shown in Fig. 4 (b). The smaller λ, the weaker the influence of the reference image.

G. Examples of Evaluation Samples

In this section, we present more evaluation samples in our method on two different test datasets: Multi-Subject bench and DreamBench bench in Fig. 6, Fig. 7, and Fig. 8.

Moreover, we present more qualitative comparison results in Fig. 9. As illustrated in the figure, our approach is more adept at focusing on the representation of distinct subjects within a single image, utilizing a query to select the necessary representation. In contrast to other methods, our method does not result in ambiguous subject extraction, a common issue in finetune-based methods. For instance, in



(a) Visual ablation results of τ.

(b) Visual ablation results of λ.

Figure 4. Visual ablation results of τ and λ.

the Dreambooth row from Fig. 9, two subjects frequently appear concurrently, indicating a low level of selectivity. When considering selectivity, generative quality, and text-image alignment, our SSR-Encoder surpasses all methods and achieves the level of finetune-based methods in terms of subject alignment.

H. Details of Our Training Data and the Multi-subject Bench

- **Details of training data.** Our model utilizes the Laion 5B dataset[19], selecting images with aesthetic scores above 6.0. The text prompts are re-captioned using BLIP2 [12]. The dataset comprises 10 million high-quality image-text pairs, with 5,000 images reserved for testing and the remainder for training. Clearly, the distribution of training data has a significant impact on our model. The more a particular type of subject data appears in the training data capt, the better our performance on that type of subject. Therefore, we further analyze the word frequency in the training data caption and report the most frequent subject descriptors in the table2.
- **Details of multi-subject bench.** The Multi-subject Bench comprises 100 images from our test data. More specifically, the data is curated based on the caption associated with each image from our test set. An image progresses to the next stage if its caption contains at least two subject descriptors. Subsequently, we verify the congruence between the caption and the image. If the image aligns with the caption and adheres to human aesthetic standards, it is shortlisted as a candidate image. Ultimately, we meticulously selected 100 images from these candidates to constitute the Multi-subject Bench.

Table 2. The most frequent subject descriptors in our training data.

Subject	frequency	Subject	frequency	subject	frequency
woman	1528518	suit	256732	dog	164819
man	1256613	trees	240771	snow	163838
people	536434	hair	229538	girl	162311
table	385643	wooden	216958	hat	157549
mountain	315765	street	212259	flowers	152308
chairs	291189	house	191785	sky	151332
dress	268058	building	168670	cat	147851

I. Compatibility with ControlNet

Our SSR-Encoder can be efficiently integrated into controllability modules. As demonstrated in Fig. 10, we present additional results of amalgamating our SSR-Encoder with ControlNet [24]. Our approach can seamlessly merge with controllability modules, thereby generating controllable images that preserve consistent character identities in alignment with reference images.

J. Compatibility with AnimateDiff

Our SSR-Encoder is not only versatile enough to adapt to various custom models and controllability modules, but it can also be effectively applied to video generation, integrating seamlessly with video generation models. In Fig. 11, we demonstrate the impact of combining our SSR-Encoder with Animatediff [5]. Despite not being trained on video data, our method can flawlessly combine with Animatediff to produce videos that maintain consistent character identities with reference images.

K. Broader Impact

Our method in subject-driven image generation holds significant potential for advancing the field of text-to-image generation, particularly in creating personalized images. This technology can be applied across various domains such as personalized advertising, artistic creation, and game design, and can enhance research at the intersection of computer vision and natural language processing. However, while the technology has numerous positive applications, it also raises ethical and legal considerations. For instance, generating personalized images using others' images without appropriate permission could infringe upon their privacy and intellectual property rights. Therefore, adherence to relevant ethical and legal guidelines is crucial. Furthermore, our model may generate biased or inappropriate content if misused. We strongly advise against using our model in user-facing applications without a thorough inspection of its output and recommend proper content moderation and regulation to prevent undesirable consequences.

L. Limitation

Due to the uneven distribution of the filtered training data, we found that the fidelity will be slightly worse for some concepts that are uncommon in our training data. This can be addressed by increasing the training data. We plan to address these limitations and extend our approach to 3D generation in our future work.

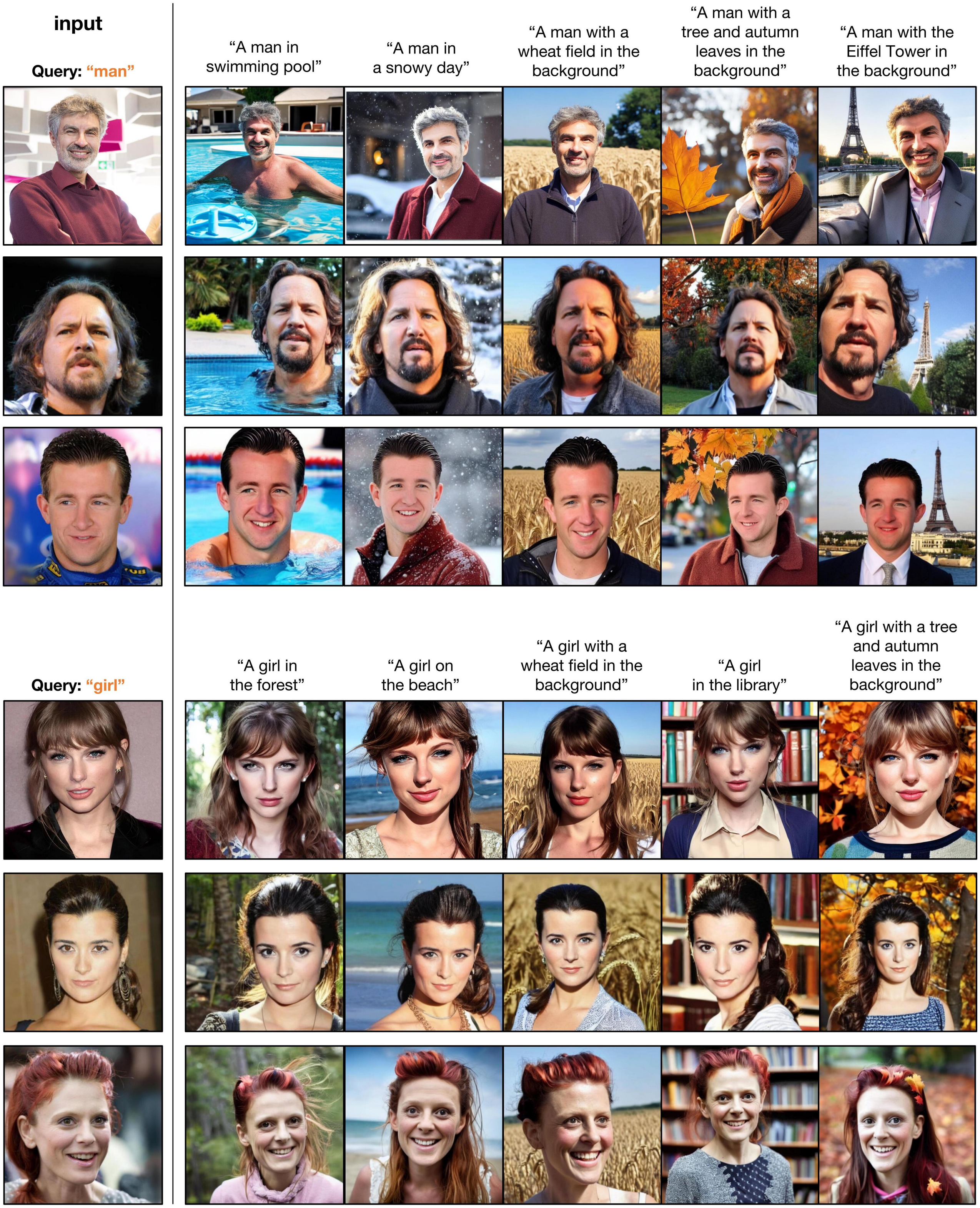

input
Query: "man"
"A man in swimming pool"
"A man in a snowy day"
"A man with a wheat field in the background"
"A man with a tree and autumn leaves in the background"
"A man with the Eiffel Tower in the background"
Query: "girl"
"A girl in the forest"
"A girl on the beach"
"A girl with a wheat field in the background"
"A girl in the library"
"A girl with a tree and autumn leaves in the background"

Figure 5. Results for human image generation.

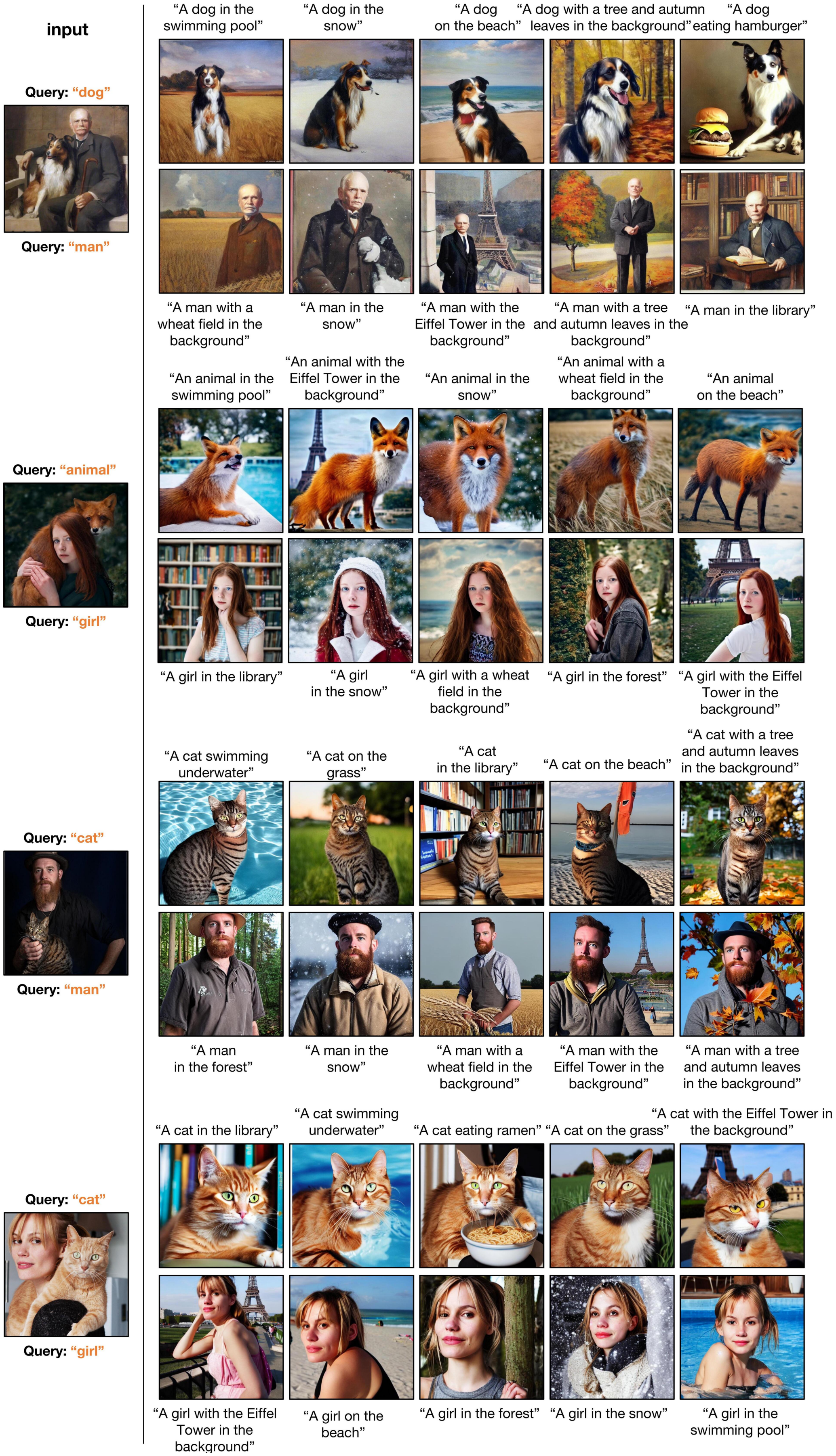

input
Query: "dog"
Query: "man"
"A dog in the swimming pool"
"A dog in the snow"
"A dog on the beach"
"A dog with a tree and autumn leaves in the background"
"A dog eating hamburger"
"A man with a wheat field in the background"
"A man in the snow"
"A man with the Eiffel Tower in the background"
"A man with a tree and autumn leaves in the background"
"A man in the library"
Query: "animal"
Query: "girl"
"An animal in the swimming pool"
"An animal with the Eiffel Tower in the background"
"An animal in the snow"
"An animal with a wheat field in the background"
"An animal on the beach"
"A girl in the library"
"A girl in the snow"
"A girl with a wheat field in the background"
"A girl in the forest"
"A girl with the Eiffel Tower in the background"
Query: "cat"
Query: "man"
"A cat swimming underwater"
"A cat on the grass"
"A cat in the library"
"A cat on the beach"
"A cat with a tree and autumn leaves in the background"
"A man in the forest"
"A man in the snow"
"A man with a wheat field in the background"
"A man with the Eiffel Tower in the background"
"A man with a tree and autumn leaves in the background"
Query: "cat"
Query: "girl"
"A cat in the library"
"A cat swimming underwater"
"A cat eating ramen"
"A cat on the grass"
"A cat with the Eiffel Tower in the background"
"A girl with the Eiffel Tower in the background"
"A girl on the beach"
"A girl in the forest"
"A girl in the snow"
"A girl in the swimming pool"

Figure 6. Examples of evaluation samples on the multi-subject bench.

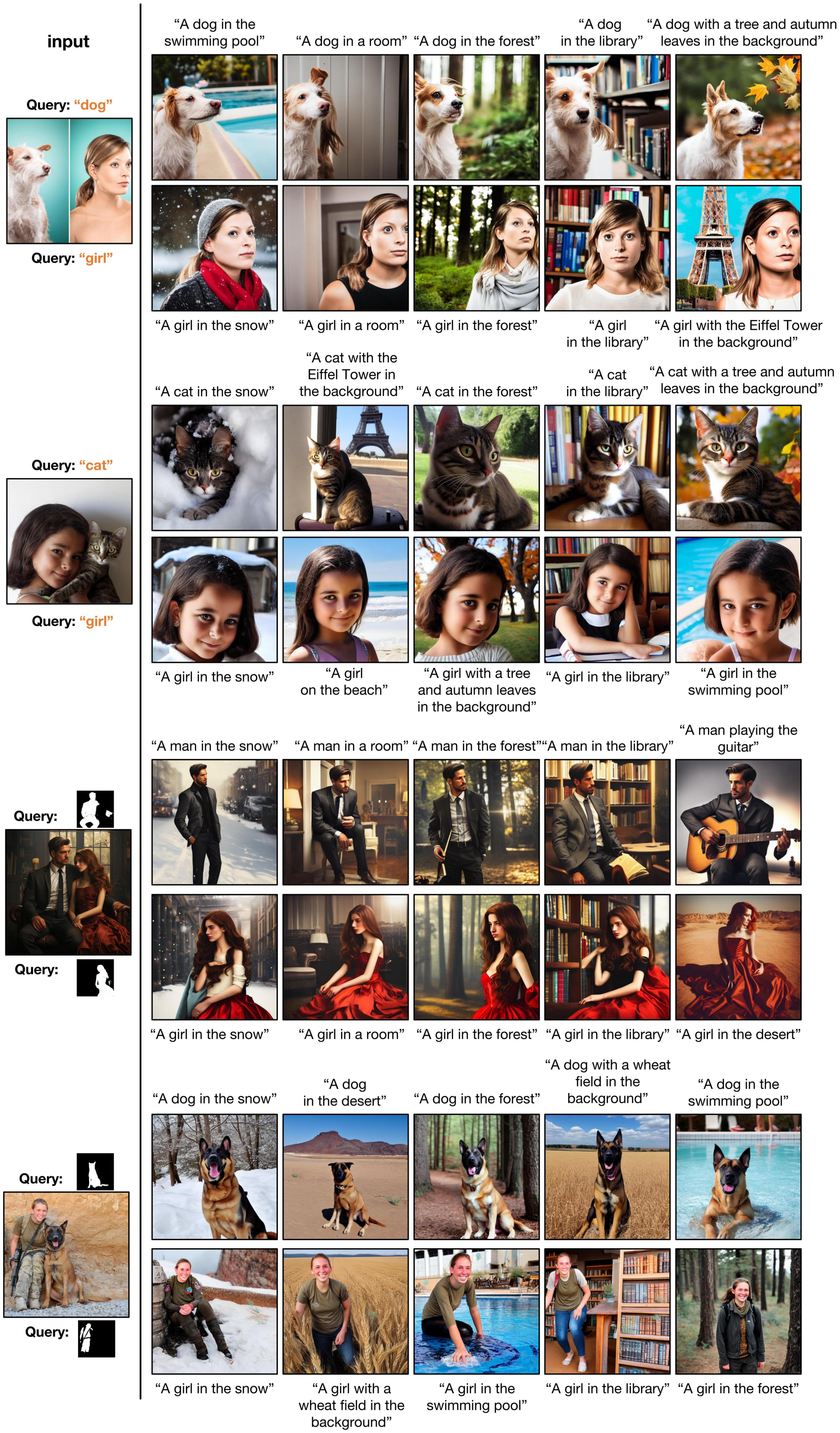


Figure 7. Examples of evaluation samples on the multi-subject bench.

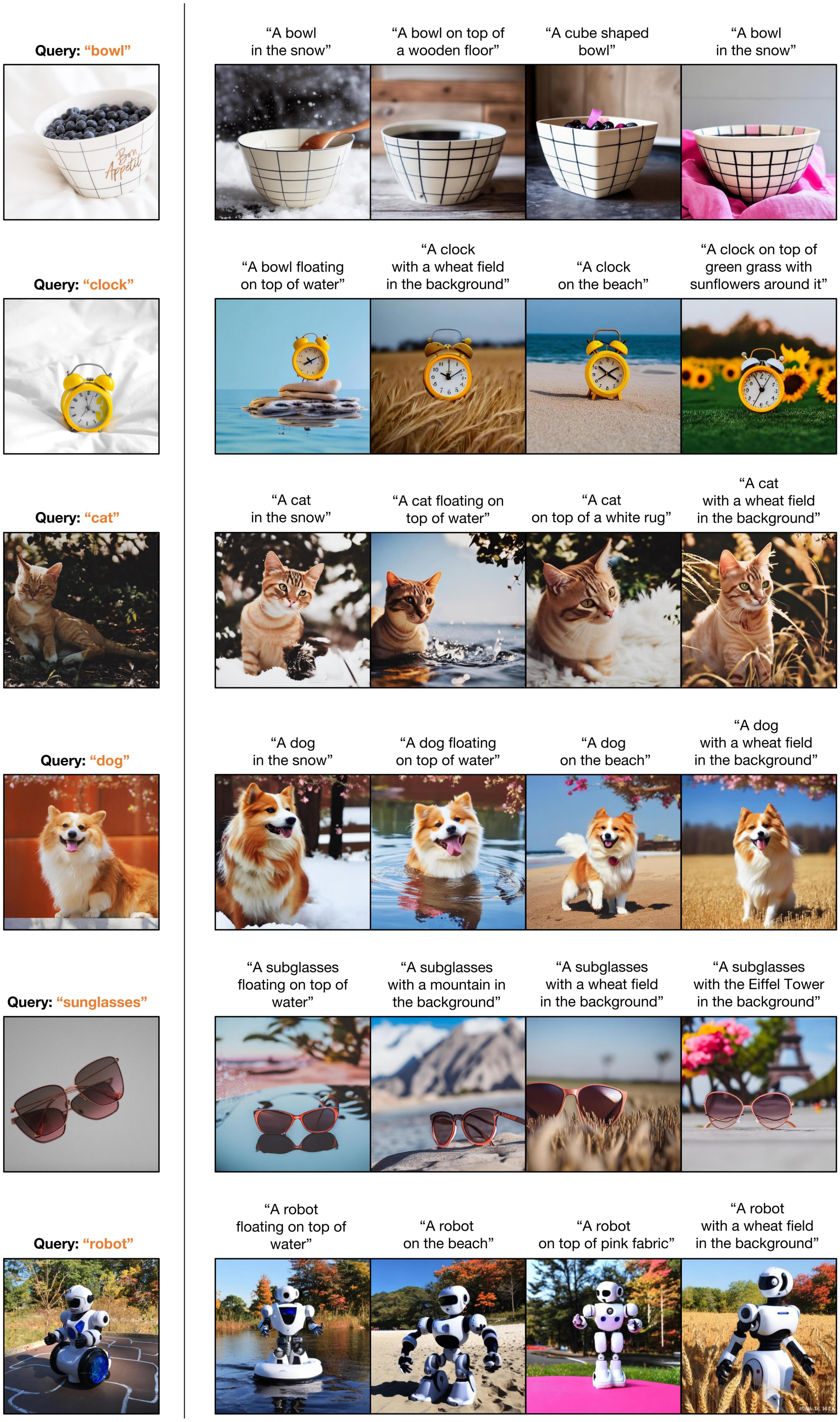

Query: "bowl"
"A bowl in the snow"
"A bowl on top of a wooden floor"
"A cube shaped bowl"
"A bowl in the snow"
Query: "clock"
"A bowl floating on top of water"
"A clock with a wheat field in the background"
"A clock on the beach"
"A clock on top of green grass with sunflowers around it"
Query: "cat"
"A cat in the snow"
"A cat floating on top of water"
"A cat on top of a white rug"
"A cat with a wheat field in the background"
Query: "dog"
"A dog in the snow"
"A dog floating on top of water"
"A dog on the beach"
"A dog with a wheat field in the background"
Query: "sunglasses"
"A subglasses floating on top of water"
"A subglasses with a mountain in the background"
"A subglasses with a wheat field in the background"
"A subglasses with the Eiffel Tower in the background"
Query: "robot"
"A robot floating on top of water"
"A robot on the beach"
"A robot on top of pink fabric"
"A robot with a wheat field in the background"

Figure 8. Examples of evaluation samples on the dreambench.

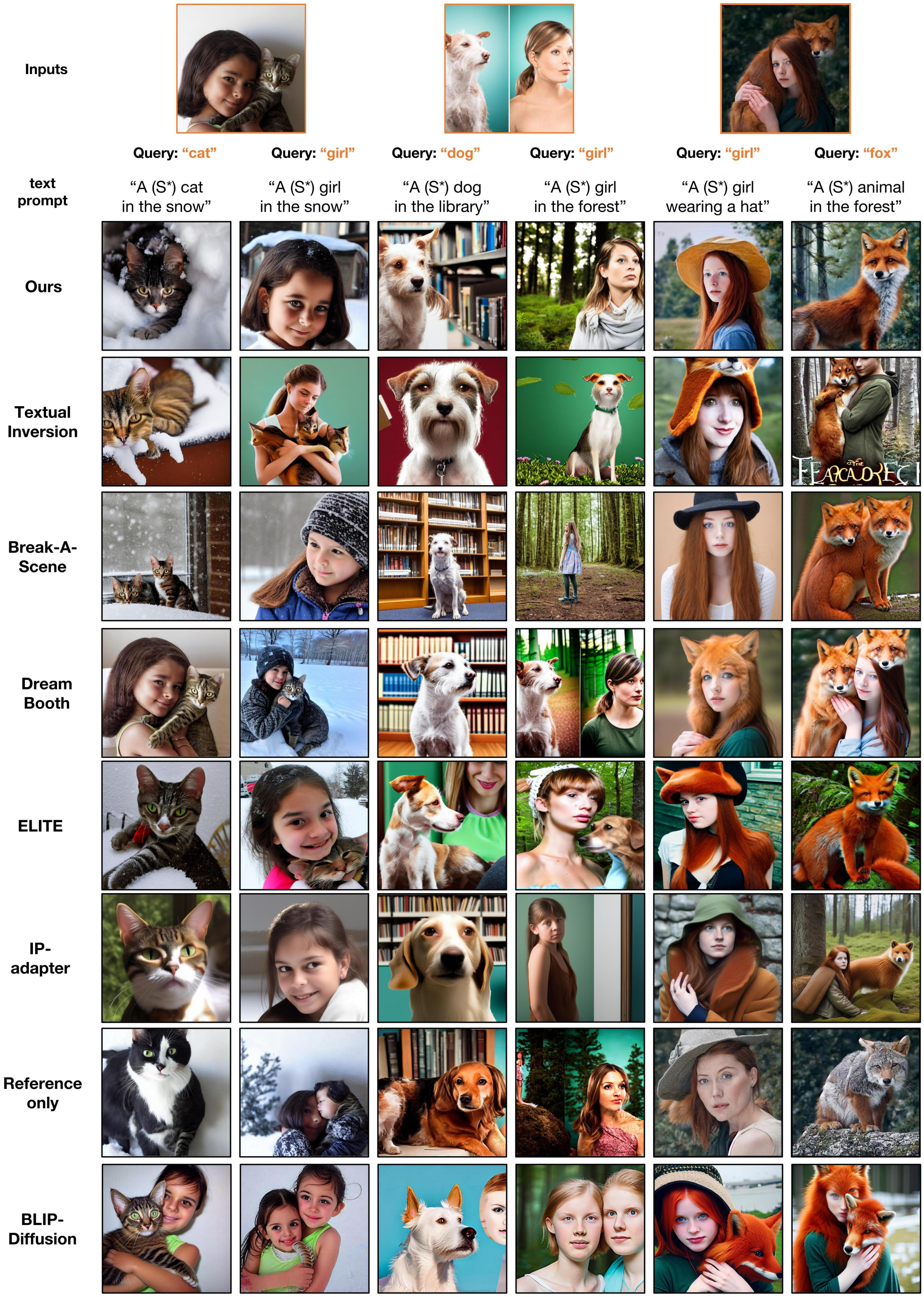


Figure 9. More results of the qualitative comparison.

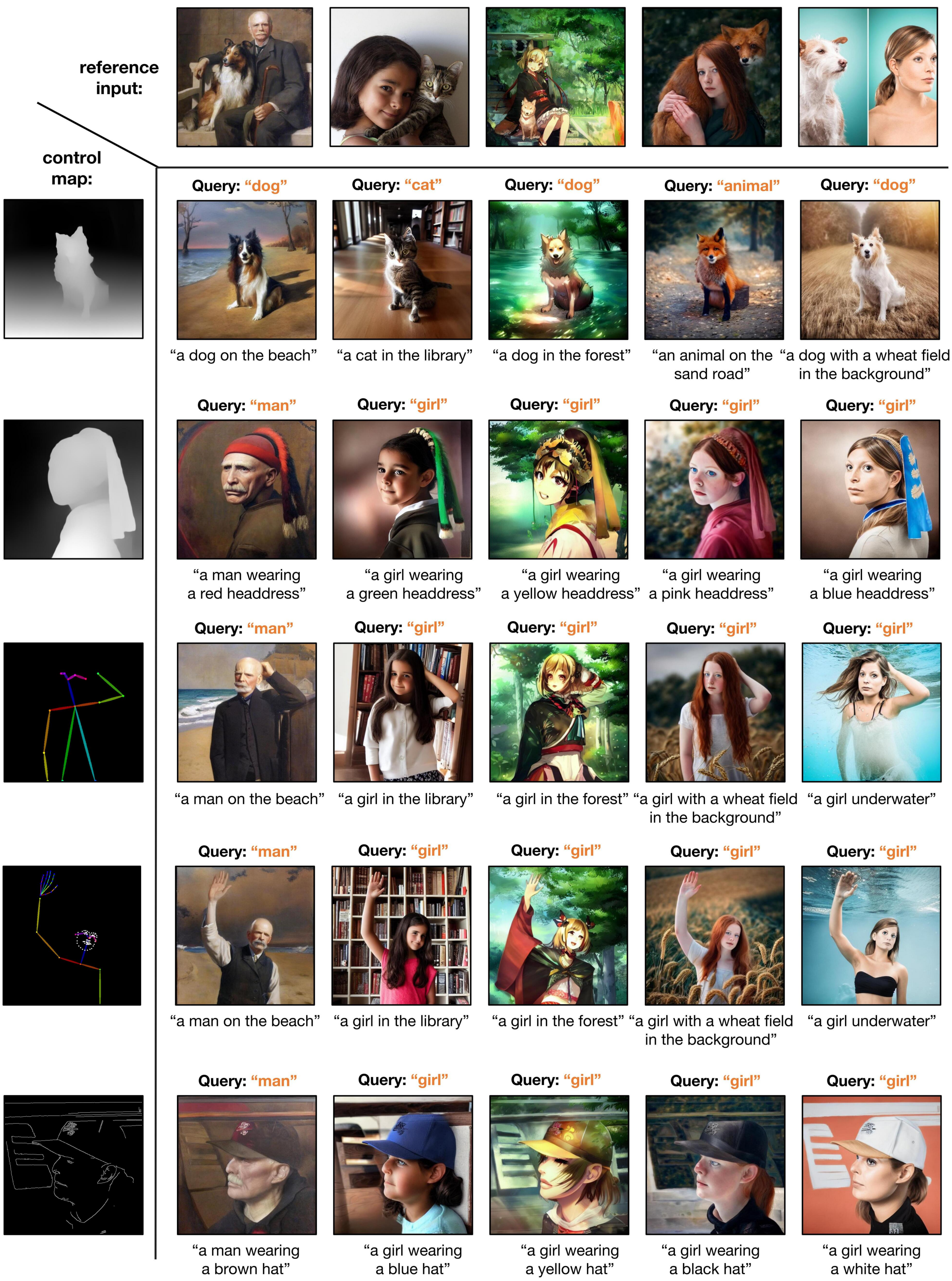


Figure 10. Results of combining our SSR-Encoder with controlnet.

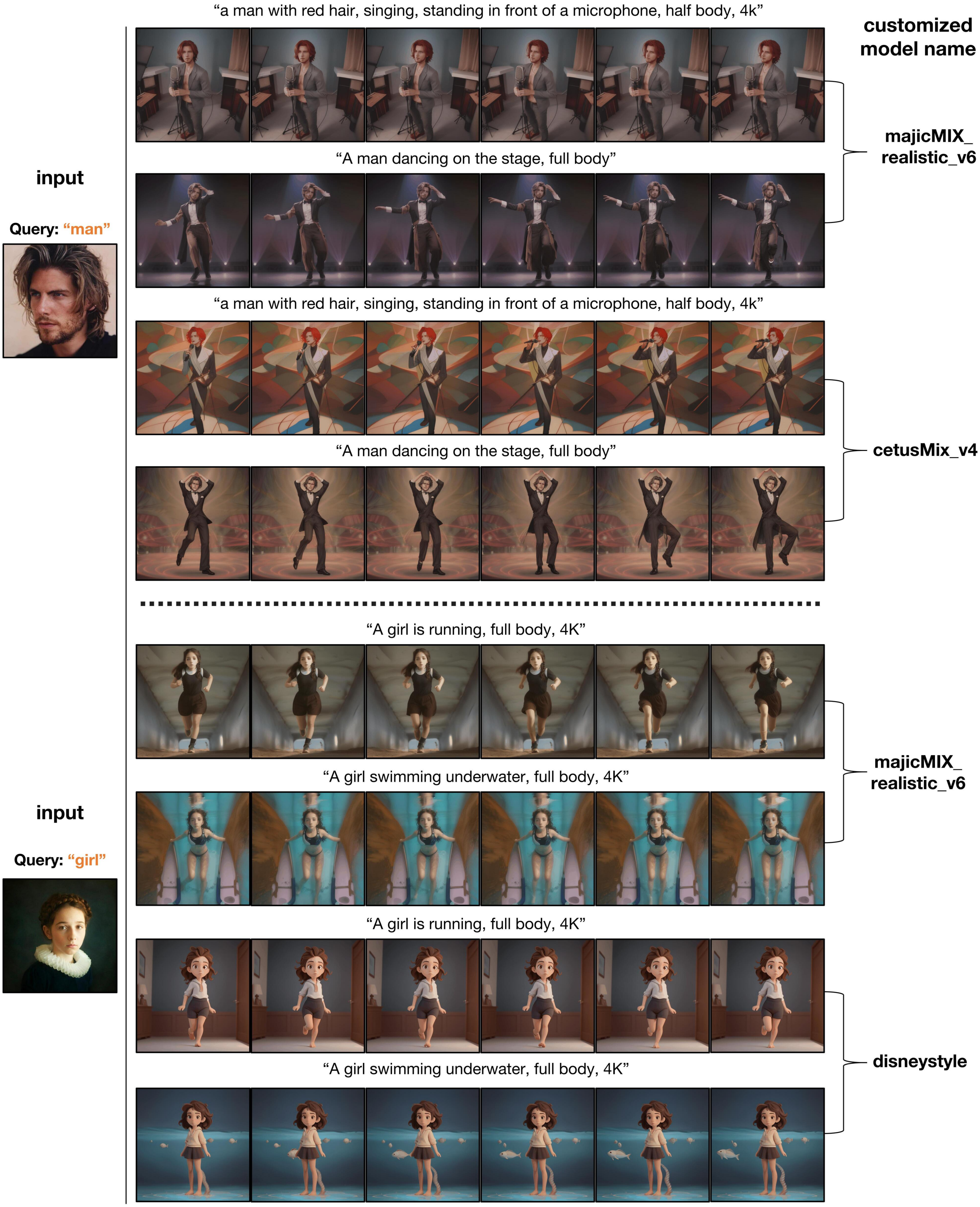

input
Query: "man"
"a man with red hair, singing, standing in front of a microphone, half body, 4k"
customized model name
majicMIX_ realistic_v6
"A man dancing on the stage, full body"
"a man with red hair, singing, standing in front of a microphone, half body, 4k"
cetusMix_v4
"A man dancing on the stage, full body"
input
Query: "girl"
"A girl is running, full body, 4K"
majicMIX_ realistic_v6
"A girl swimming underwater, full body, 4K"
"A girl is running, full body, 4K"
disneystyle
"A girl swimming underwater, full body, 4K"

Figure 11. Results of combining our SSR-Encoder with Animatediff.

References

[1] 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

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 1

[4] 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, 2022. 2

[5] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023. 1, 4

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. 1

[8] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 2

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 2

[10] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128(7): 1956–1981, 2020. 3

[11] 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. 3

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023. 4

[13] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410, 2023. 3

[14] Mikubill. sd-webui-controlnet, 2023. GitHub repository. 2

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[18] 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. 2022. 2, 3

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 1, 2, 4

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020. 1

[21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020. 1

[22] 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

[23] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 3

[24] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023. 1, 4

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018. 2