MagiCapture: High-Resolution Multi-Concept Portrait Customization

Junha Hyung[*1], **Jaeyo Shin**[*2], **and Jaegul Choo**[1]
[1] KAIST AI, [2] Sogang University
{sharpeeee, jchoo}@kaist.ac.kr, tlswody123@sogang.ac.kr

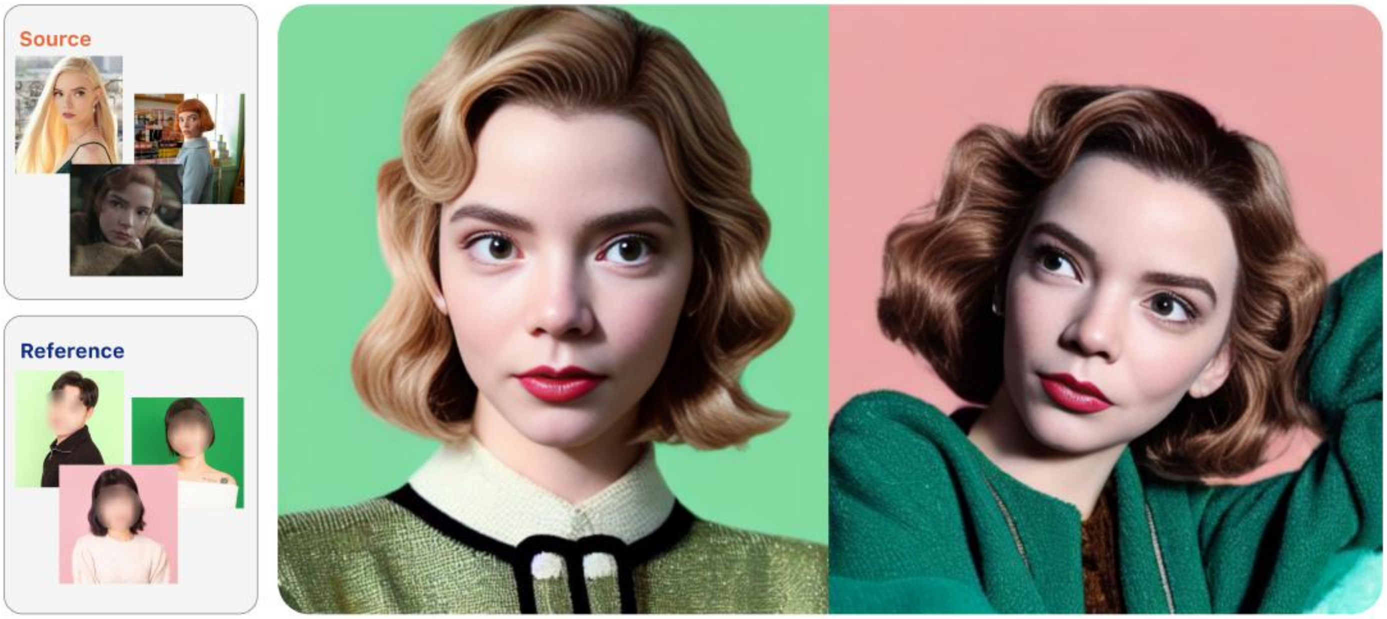


Figure 1: Generated results of the proposed **MagiCapture**, a multi-concept personalization method for integrating subject and style concepts to generate high-resolution portrait images using just a few subject and style references.

Abstract

Large-scale text-to-image models including Stable Diffusion are capable of generating high-fidelity photorealistic portrait images. There is an active research area dedicated to personalizing these models, aiming to synthesize specific subjects or styles using provided sets of reference images. However, despite the plausible results from these personalization methods, they tend to produce images that often fall short of realism and are not yet on a commercially viable level. This is particularly noticeable in portrait image generation, where any unnatural artifact in human faces is easily discernible due to our inherent human bias. To address this, we introduce MagiCapture, a personalization method for integrating subject and style concepts to generate high-resolution portrait images using just a few subject and style references. For instance, given a handful of random selfies, our fine-tuned model can generate high-quality portrait images in specific styles, such as passport or profile photos. The main challenge with this task is the absence of ground truth for the composed concepts, leading to a reduction in the quality of the final output and an identity shift of the source subject. To address these issues, we present a novel Attention Refocusing loss coupled with auxiliary priors, both of which facilitate robust learning within this weakly supervised learning setting. Our pipeline also includes additional post-processing steps to ensure the creation of highly realistic outputs. MagiCapture outperforms other baselines in both quantitative and qualitative evaluations and can also be generalized to other non-human objects.

Introduction

To obtain high-quality portrait images suitable for resumes or wedding events, individuals typically have to visit a photo studio, followed by a costly and time-consuming process of photo retouching. Imagine a scenario where all that's required is a few selfie images and reference photos, and you could receive high-quality portrait images in specific styles,

[*] These authors contributed equally.

such as passport or profile photos. This paper aims to automate this process.

Recent advancements in large-scale text-to-image models, such as Stable Diffusion [21] and Imagen [23], have made it possible to generate high-fidelity, photorealistic portrait images. The active area of research dedicated to personalizing these models seeks to synthesize specific subjects or styles using provided sets of train images. In this work, we formulate our task as a multi-concept customization problem. Here, the source content and reference style are learned respectively, and the composed output is generated. Unlike text-driven editing, using reference images allows users to provide fine-grained guidance, making it more suitable for this task.

However, despite the promising results achieved by previous personalization methods, they often produce images that lack realism and fall short of commercial viability. This problem primarily arises from attempting to update the parameters of large models using only a small number of images. This decline in quality becomes even more evident in a multi-concept generation, where the absence of ground truth images for the composed concepts frequently leads to the unnatural blending of disparate concepts or deviation from the original concepts. This issue is particularly conspicuous in portrait image generation, as any unnatural artifacts or shifts in identity are easily noticeable due to our inherent human bias.

To address these issues, we present MagiCapture, a multi-concept personalization method for the fusion of subject and style concepts to generate high-resolution portrait images with only a few subject and style references. Our method employs composed prompt learning, incorporating the composed prompt as part of the training process, which enhances the robust integration of source content and reference style. This is achieved through the use of pseudo labels and auxiliary loss. Moreover, we propose the Attention Refocusing loss in conjunction with a masked reconstruction objective, a crucial strategy for achieving information disentanglement and preventing information leakage during inference. MagiCapture outperforms other baselines in both quantitative and qualitative assessments and can be generalized to other non-human objects with just a few modifications.

The main contributions of our paper are as follows:

- We introduce a multi-concept personalization method capable of generating high-resolution portrait images that faithfully capture the characteristics of both source and reference images.
- We present a novel Attention Refocusing loss combined with masked reconstruction objective, effectively disentangling the desired information from input images and preventing information leakage during the generation process.
- We put forth a composed prompt learning approach that leverages pseudo-labels and auxiliary loss, facilitating the robust integration of source content and reference style.
- In both quantitative and qualitative assessments, our method surpasses other baseline approaches and, with minor adjustments, can be adapted to generate images of non-human objects.

Related Work

Text-to-image diffusion models Diffusion models [10, 27, 28, 26] have recently achieved remarkable success in image generation, driving advancements in various applications and fields. Their powerful performance has significantly propelled the field of text-guided image synthesis [16, 12, 23, 19] forward. In particular, large-scale text-to-image diffusion models, trained on extensive text-image pair datasets, have set new benchmarks. Notable examples include Stable diffusion [30] and Imagen [23]. Our work is built upon the pre-trained stable diffusion model.

Personalization of Text-to-image Models. Personalizing generative models for specific concepts is a key goal in the vision field. With the rise of GANs, there have been efforts to fine-tune GANs, like Pivotal Tuning [20], based on GAN inversion [36].

More recently, studies have sought to personalize diffusion models using small image datasets, typically $3 \sim 5$ images, associated with a particular object or style and incorporating specialized text tokens to embed such concepts. For instance, when customizing models for a specific dog, the prompt "a $[V1]$ dog" is used so that the special token can learn information specific to the dog. DreamBooth [22] fine-tunes entire weights, Textual Inversion [6] adjusts text embeddings, and Custom Diffusion [14] adapts the mapping matrix for the cross-attention layer. While effective in learning concepts, these models sometimes generate less realistic or identity-losing images. Methods like ELITE [32] and InstantBooth [25] employ a data-driven approach for encoder-based domain tuning, which is not directly comparable to our approach.

Our method differs from concurrent works like SVDiff [8], FastComposer [33], and Break-A-Scene [1], which use similar techniques like attention loss or composed prompts. Unlike SVDiff's collage approach (Cut-Mix-Unmix), our method is tailored for style-mixed outputs, enhancing the quality of multi-concept portraits. Distinct from FastComposer and Break-A-Scene, our attention loss only targets regions in the attention map not present in the ground-truth mask ($A_k[i, j]$ for all $(i, j) \in \{(i, j)|M_v[i, j] = 0\}$), allowing for the varying optimal values for other areas.

Preliminaries

Diffusion Models. Diffusion models [10, 27, 28, 26] are a class of generative models that create images through an iterative denoising process. These models comprise a forward and backward pass. During the forward pass, an input image $x^{(0)}$ is progressively noised using the equation $x^{(t)} = \sqrt{\alpha_t}x^{(0)} + \sqrt{1 - \alpha_t}\epsilon$, where ϵ represents standard Guassian noise and $\{\alpha_t\}$ is a pre-defined noise schedule with timestep t, $1 < t < T$. During backward pass, the generated image is obtained by denoising the starting noise x_T

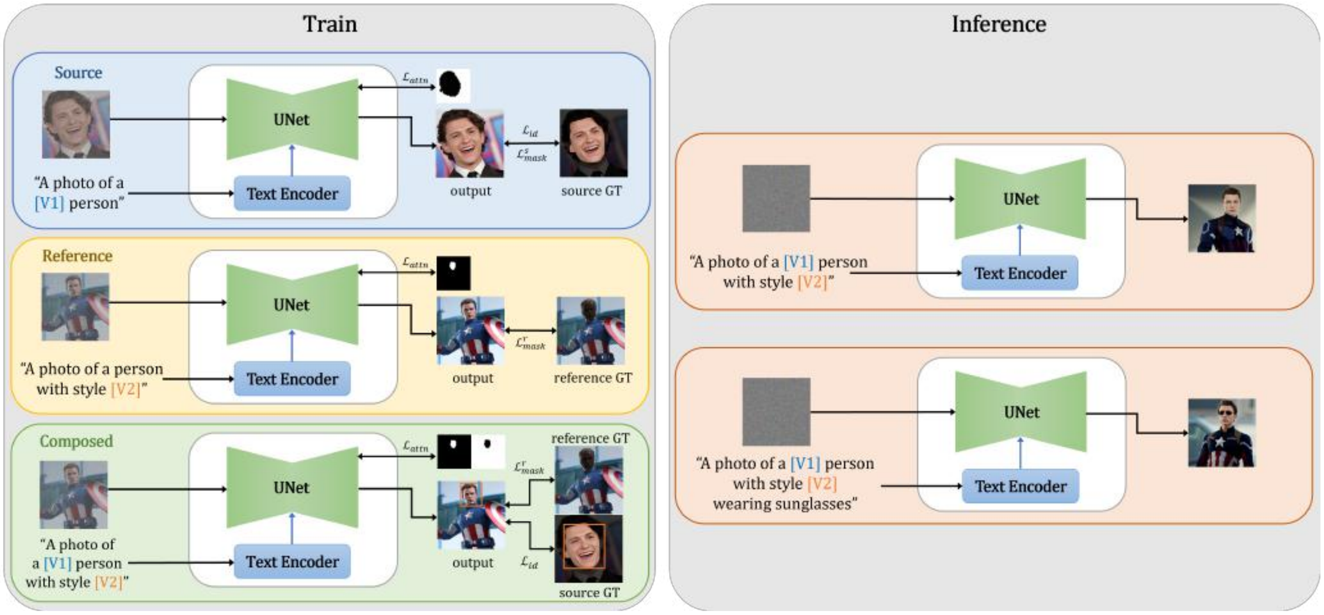


Figure 2: The overall pipeline of **MagiCapture**, where the training process is formulated as multi-task learning of three different tasks: source, reference, and composed prompt learning. In the composed prompt learning, reference style images serve as pseudo-labels, along with auxiliary identity loss between the source and predicted images. Attention Refocusing loss is applied to all three tasks. After training, users can generate high-fidelity images with integrated concepts and can further manipulate them using varying text conditions.

using a UNet $\epsilon_\theta(x^{(t)}, t)$, which is trained to predict noise at the input timestep t. Latent diffusion models (LDM) [21] are a variant of diffusion models where the denoising process occurs in the latent space. Specifically, an image encoder $\mathcal{E}$ is used to transform the input image x into a latent representation z, such that $\mathcal{E}(x) = z$. During inference, the denoised latent representation is decoded to produce the final image $x^{(0)\prime} = \mathcal{D}(z^{(0)})$, where $\mathcal{D}$ represents the decoder of an autoencoder. Stable diffusion [30] is a text-guided latent diffusion model (LDM) trained on large-scale text-image pairs. It has the following objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,c,\epsilon,t}\Big[||\epsilon_\theta(z^{(t)}, t, c) - \epsilon||_2^2\Big], \tag{1}$$

where c refers to the text condition.

Attention maps Large-scale text-to-image diffusion models utilize cross-attention layers for text-conditioning. In Stable Diffusion [21], CLIP text encoder [18] is used to produce text embedding features. These text embeddings are then transformed to obtain the key K and value V for the cross-attention layer through linear mapping, and spatial feature of image is projected to query Q. The attention map of the cross-attention layer is computed as:

$$A = \text{softmax}\Big(\frac{QK^T}{\sqrt{d}}\Big). \tag{2}$$

The attention map corresponding to a specific token with index k can be obtained as $A_k = A[k]$. Such attention maps are useful for visualizing the influence of individual tokens in the text prompt. Moreover, they can be altered or manipulated for the purpose of image editing, as demonstrated in Prompt-to-Prompt [9].

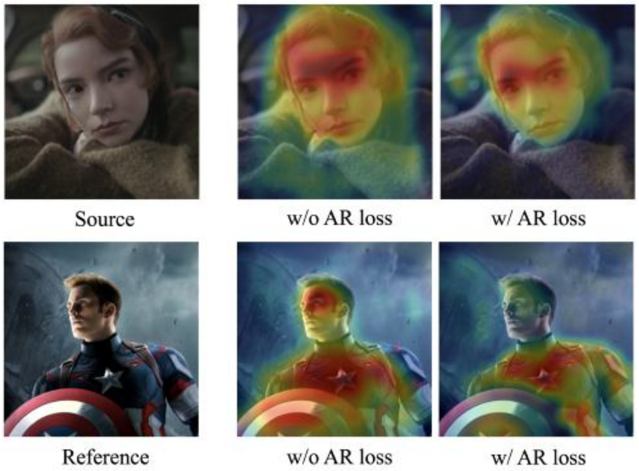


Figure 3: Visualization of aggregated attention maps from UNet layers before and after the application of Attention Refocusing (AR) loss illustrates its importance in achieving information disentanglement and preventing information spill.

Method

Given a small set of source images and reference style images, the goal of this paper is to synthesize images that integrate the source content with the reference style. While our method is primarily designed for generating portrait images, it can be easily adapted to handle other types of content with minor modifications. We utilize the customization of each concepts during the optimization phase and employ a composed prompt during inference to generate multi-concept images. A comprehensive overview of our approach is depicted in Fig. 2, and the details of our method will be elaborated upon in the subsequent sections.

Two-phase Optimization. Similar to Pivotal Tuning [20] in GAN inversion, our method consists of two-phase optimization. In the first phase, we optimize the text embeddings for the special tokens $[V^*]$ using the reconstruction objective as in [6]. While optimizing the text embeddings is not sufficient for achieving high-fidelity customization, it serves as a useful initialization for the subsequent phase. In the second phase, we jointly optimize the text embeddings and model parameters with the same objective. Rather than optimizing the entire model, we apply the LoRA [11], where only the residuals ΔW of the projection layers in the cross-attention module are trained using low-rank decomposition. Specifically, the updated parameters are expressed as:

$$W^{'} = W + \Delta W,\ \Delta W = UV^T, \tag{3}$$

where $U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}$, and $r << n, m$. Empirically, we find that this two-phase optimization coupled with LoRA strikes a favorable balance between reconstruction and generalization. It preserves the model's generalization capabilities for unseen prompts while effectively capturing the finer details of the source images.

Masked Reconstruction. In our approach, a source prompt c_s (e.g., A photo of $[V1]$ person.) and a reference prompt c_r (e.g., A photo of a person in the $[V2]$ style.) are used to reconstruct the source image I_s and a target style image I_r respectively. It is crucial to disentangle the identity of the source subject from non-facial regions, such as the background and clothing, to prevent this unwanted information from being encoded into the special token $[V1]$. Similarly, we need to disentangle the reference image to ensure that the facial details of the person in the reference image are not embedded into the special token $[V2]$. To achieve this, we propose to use a masked reconstruction loss. Specifically, we employ a mask that indicates the relevant region and apply it element-wise to both the ground truth latent code and the predicted latent code. In the context of portrait generation, a source mask M_s indicates the facial region of the image I_s, and a target mask M_r denotes the non-facial areas of the reference image I_r. Formally, the masked reconstruction loss for the source and the reference prompts are given by:

$$\mathcal{L}^s_{mask} = \mathbb{E}_{z_s,c_s,\epsilon,t}\Big[||\epsilon \odot M_s - \epsilon_\theta(z_s^{(t)}, t, c_s) \odot M_s||_2^2\Big], \tag{4}$$

$$\mathcal{L}^r_{mask} = \mathbb{E}_{z_r,c_r,\epsilon,t}\Big[||\epsilon \odot M_r - \epsilon_\theta(z_r^{(t)}, t, c_r) \odot M_r||_2^2\Big], \tag{5}$$

Method	CSIM ↑	Style ↑	Aesthetic ↑
DreamBooth	0.102	0.720	5.770
Textual Inversion	0.224	0.623	5.670
Custom Diffusion	0.436	0.606	5.263
Ours w/o AR & CP	0.429	0.726	6.178
Ours	**0.566**	**0.730**	**6.218**

Table 1: Quantitative comparison of our method against DreamBooth [22], Textual Inversion [6], and Custom Diffusion [14]. Our method outperforms other baselines in terms of identity similarity measured between the source images (**CSIM**), masked CLIP similarity measure (**Style**), and **Aesthetic score** [24].

where $z_s^{(t)}$ and $z_r^{(t)}$ are the source and reference noised latent at timestep $t \sim$ Uniform(1, T) and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Composed Prompt Learning. Generating images with a composed prompt c_c such as "A photo of a $[V1]$ person in the $[V2]$ style," leads to undefined behavior because the model had not been customized on such prompts. Typically, the resulting images generated using these unseen composed prompts suffer from a shift in the identity of the source subject and a decline in output quality. To address this issue, we include training on the composed prompt. However, no ground truth image exists for such a prompt. We approach this challenge as a weakly-supervised learning problem, where there are no available ground truth labels. We craft pseudo-labels and develop an auxiliary objective function to suit our needs. In the context of the portrait generation task, we want to retain the overall composition, pose, and appearance from the reference style image, excluding the facial identity. To achieve this, we employ the masked reconstruction objective given by:

$$\mathcal{L}^c_{mask} = \mathbb{E}_{z_r,c_c,\epsilon,t}\Big[||\epsilon \odot M_r - \epsilon_\theta(z_r^{(t)}, t, c_c) \odot M_r||_2^2\Big]. \tag{6}$$

For the facial regions, we use an auxiliary identity loss that utilizes a pre-trained face recognition model [5] $\mathcal{R}$ and cropping function $\mathcal{B}$ conditioned by the face detection model [4]:

$$\mathcal{L}_{id} = \mathbb{E}_{\hat{x}^{(0)},I_s}\Big[1 - \cos(\mathcal{R}(\mathcal{B}(\hat{x}^{(0)})), \mathcal{R}(\mathcal{B}((I_s)))\Big], \tag{7}$$

where cos denotes the cosine similarity and $\hat{x}^{(0)} = \mathcal{D}(\hat{z}^{(0)})$ refers to the estimated clean image from $z_r^{(t_{id})}$ using Tweedie's formula [13]. Timestep t_{id} is sampled as $t_{id} \sim$ Uniform(1, $T^{'}$), where $T^{'} < T$, to avoid blurry and inaccurate $\hat{x}^{(0)}$ estimated from noisy latent with large timesteps, which can impair cropping or yield odd facial embeddings.

We augment the composed prompt c_c by randomly selecting from predefined prompt templates to boost editing stability and generalization.

Attention Refocusing. When optimizing with training images, it is vital to achieve ***information disentanglement***, ensuring that special tokens exclusively embed the information of the region of interest, denoted as M_v for $v \in \{s, r\}$. However, the masked reconstruction objective falls short of this goal because the presence of transformer layers in the

UNet backbone gives the model a global receptive field. The same limitation applies to denoising steps in the inference stage, where we desire attention maps of special tokens to focus only on the intended areas. For instance, in the portrait generation task, the special token $[V1]$ should only attend to facial regions when generating images to avoid ***information spill***. We observe that information spill is more prevalent when the model encounters an unseen prompt during inference. Fig. 3 demonstrates that special tokens do indeed attend to unwanted regions.

To solve this issue, we propose a novel Attention Refocusing (AR) loss, which steers the cross attention maps A_k of the special token $[V^*]$ (where $k = \text{index}([V^*])$) using a binary target mask. Our AR loss incorporates two crucial details: First, it is applied only to regions where $\neg M_v$, where the mask value is zero. For the attention map values $A_k[i,j]$ where $(i,j) \in \{(i,j)|M_v[i,j] = 1\}$, the optimal values can vary across different UNet layers and denoising time steps, so they do not necessarily have to be close to 1. Conversely, for $A_k[i,j]$ where $(i,j) \in \{(i,j)|M_v[i,j] = 0\}$, the values should be forced to 0 to achieve information disentanglement during training and minimize information spill in the inference stage. Second, it is essential to scale the attention maps to the [0,1] range. Both of these techniques are required to avoid disrupting the pre-trained transformer layers' internal operations, which would lead to corrupted outputs. The Attention Refocusing loss can be formulated as follows:

$$\mathcal{L}_{attn} = \mathbb{E}_{k,v\in\{s,r\}}\Big[||(\mathcal{S}(A_k) - M_v)\odot\neg M_v||_2^2\Big], \qquad (8)$$

where $\mathcal{S}(\cdot)$ refers to a scaling function.

Postprocessing. The quality of images generated in a few-shot customization task is typically constrained by the capabilities of the pretrained text-to-image model used. Moreover, when provided with low-resolution source and target images, the fine-tuned model tends to produce lower-quality images. To overcome these limitations and further enhance the fidelity of the generated images, our pipeline includes optional postprocessing steps. Specifically, we employ a pre-trained super-resolution model [31] and a face restoration model [35] to further improve the quality of the generated samples.

Experiments

Training Details. Our method utilizes pre-trained Stable Diffusion V1.5 [21]. The first training phase consists of a total of 1200 steps, with a learning rate 5e-4 for updating the text embeddings. In the second LoRA phase, the learning rate is 1e-4 for the projection layers and 1e-5 for the text embeddings, with a total of 1500 training steps. The model is trained on a single GeForce RTX 3090 GPU, using a batch size of 1 and gradient accumulation over 4 steps. For all experiments, we employ 4 to 6 images for both the source and reference images. Please refer to the supplement for more details.

Comparisons. The results of our method are demonstrated in Fig. 4. We compare our method with other personalization methods including DreamBooth [22], Textual Inversion [6], and Custom Diffusion [14] using the same source and reference images. We choose 10 identities, 7 from VGGFace [2] and 3 in-the-wild identities gathered from the internet. We also manually select 10 style concepts, leading to 100 id-style pairs. For each pair, we train each baseline and our model, then generate 100 images with the *composed prompt* for each of the trained model, resulting in 10,000 samples per baseline. Qualitative comparisons are shown in Fig. 5, where our method outperforms other baselines in image fidelity and source-reference image reflection.

Method	ID ↑	Style ↑	Fidelity ↑
DreamBooth	2.025	3.648	2.683
Textual Inversion	2.907	3.038	2.965
Custom Diffusion	3.223	2.260	2.980
Ours	**4.055**	**4.165**	**4.293**

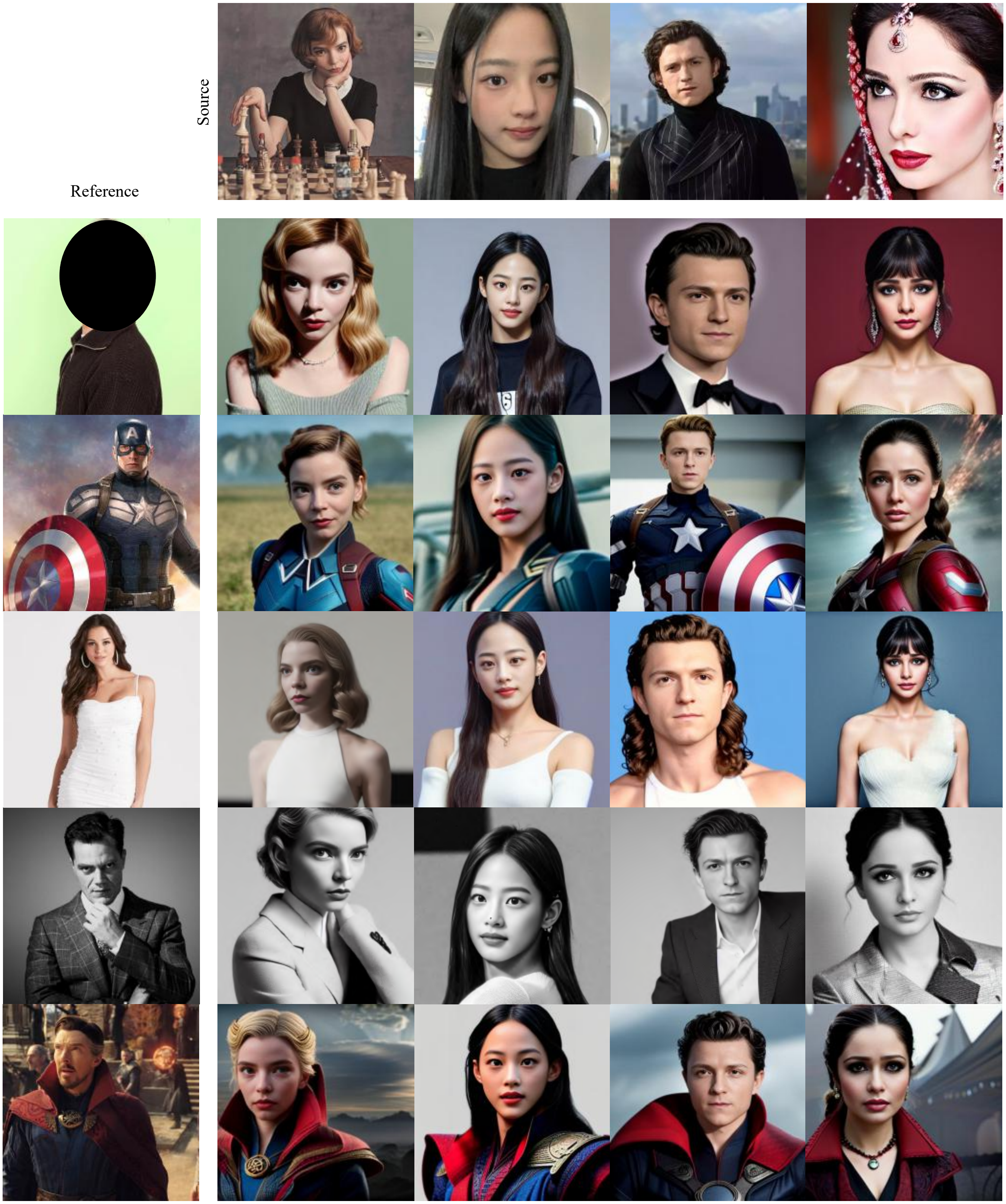
Table 2: User study of our method against DreamBooth [22], Textual Inversion [6], and Custom Diffusion [14]. Our method outperforms other baselines in terms of identity similarity score (**ID**), style similarity measure (**Style**), and image fidelity score (**Fidelity**).

We assess the facial appearance similarity between the source and generated portrait images by measuring the cosine similarity between their facial embeddings, using a pre-trained recognition network (CSIM) [34].

Another important aspect of evaluation is style preservation, where we measure how well the results replicate the style of the reference images. We compute the cosine similarity between the *masked* CLIP [18] image embeddings of the reference and generated images, where facial regions are masked to exclude facial appearance from the assessment. We use CLIP similarity instead of texture similarity [7] since the term *style* in our paper encompasses broader concepts such as image geometry and composition, in addition to texture and appearance of non-facial regions. Finally, we evaluate the overall image fidelity with the LAION aesthetic predictor [24]. Table 1 shows that our method outperforms other baselines in all three metrics. Additionally, we conduct a user study involving 30 participants who were asked to rate images for ID preservation, style preservation, and image fidelity on a 1-5 scale. Table 2 summarizes the results, with our method consistently scoring higher than other baselines.

We observed that DreamBooth often overfits to the reference style images, leading to high style scores but low CSIM scores. Conversely, Textual Inversion tends to underfit both the source and reference images, resulting in low-fidelity images that fail to preserve appearance details. Custom Diffusion better preserves source identity compared to the others, but still cannot consistently perform well for the composed prompt, leading to identity shifts and unnatural images.

Ablation Study. As shown in Fig. 3, we find that Attention Refocusing loss effectively prevents attention maps from attending to unwanted regions, mitigating information spill and promoting information disentanglement. Empirically, we observe that the Attention Refocusing loss should only be applied during the second phase of training (LoRA



[redacted]

Figure 4: Curated results of MagiCapture.

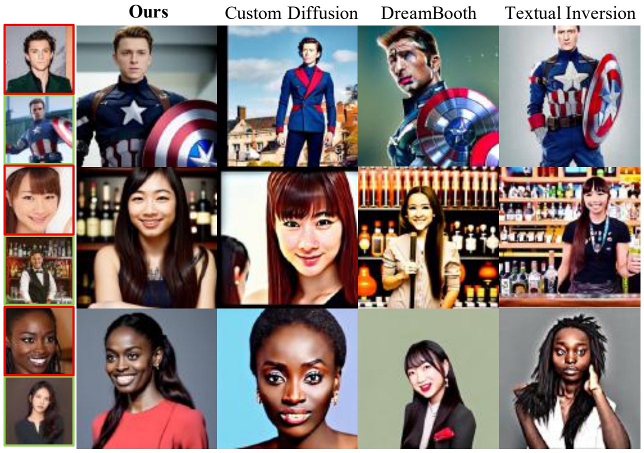


Figure 5: Qualitative comparisons of MagiCapture with other baseline methods.

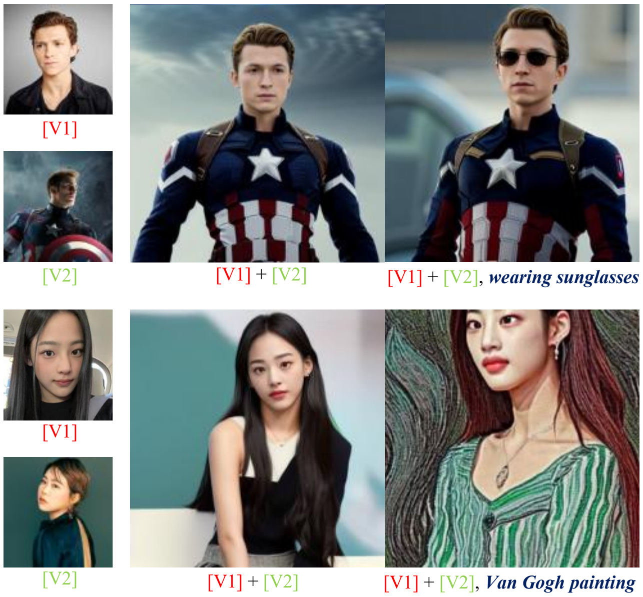

[V1]
[V2]
[V1] + [V2]
[V1] + [V2], wearing sunglasses
[V1]
[V2]
[V1] + [V2]
[V1] + [V2], Van Gogh painting

Figure 6: Users can further manipulate the composed results using prompts with additional description.

training). We infer that text embeddings are not well-suited for learning geometric information related to attention maps. Moreover, without composed prompt learning, the generated images often exhibit undefined behaviors where only one of the source or reference sets is evident in the image, without blending. We present the evaluation metrics for both the presence and absence of composed prompt learning (CP) and Attention Refocusing (AR) in Table 1. For more results and detailed analysis, please refer to the supplement.

Applications. Since our method is robust to generalizations, users can further manipulate the composed results using prompts with more descriptions (e.g., $c_c^{'}$ = "A photo of $[V1]$ person in the $[V2]$ style, wearing sunglasses."). We demonstrate such results in Fig. 6 and in the supplement. Furthermore, our method is adaptable for handling different types of content, including non-human images. For methodologies and results related to non-human content, please refer to the supplementary material.

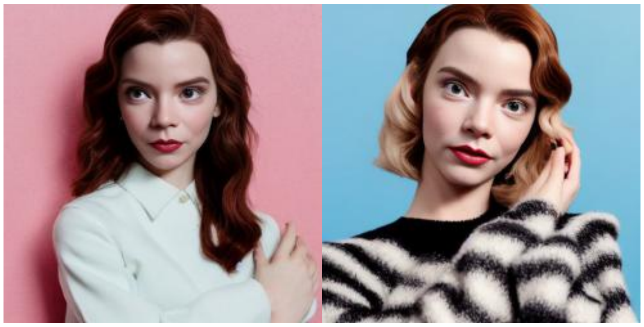


Figure 7: Failure cases: Proposed method occasionally produces abnormal body parts such as limbs, fingers

Limitations and Conclusions

Our method occasionally produces abnormal body parts such as limbs, fingers, as shown in Fig. 7. Furthermore, the model tends to exhibit lower fidelity for non-white subjects and demonstrates a noticeable gender bias—for instance, it struggles to accurately generate images of men wearing wedding dresses. These issues are largely related to the inherent biases of the pre-trained text-to-image models, and addressing these problems within a few-shot setting represents a significant avenue for future research. We acknowledge the ethical implications of our work and are committed to taking them seriously. We are also proactive in leading and supporting efforts to prevent potential misuse of our contributions.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

[1] Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *arXiv preprint arXiv:2305.16311*.

[2] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

[3] Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

[4] Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

[5] Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

[6] Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

[7] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

[8] Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*.

[9] Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control.

[10] Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

[11] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[12] Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

[13] Kim, K.; and Ye, J. C. 2021. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34: 864–874.

[14] Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

[15] Kwon, G.; and Ye, J. C. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*.

[16] Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

[17] Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345. Springer.

[18] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

[19] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

[20] Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1): 1–13.

[21] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

[22] Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

[23] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

[24] Schuhmann, C. Aug 2022. Laion aesthetics.

[25] Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

[26] Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

[27] Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

[28] Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

[29] Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

[30] von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

[31] Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*.

[32] Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

[33] Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431*.

[34] Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9459–9468.

[35] Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. In *NeurIPS*.

[36] Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. In *European conference on computer vision*, 592–608. Springer.

Supplementry Materials

Training Details

MagiCapture The loss function for the first phase training is given as:

$$\mathcal{L}^s_{mask} + \mathcal{L}^r_{mask}. \tag{9}$$

For the second phase LoRA training, composed prompt learning and Attention Refocusing loss is added:

$$\mathcal{L}^s_{mask} + \mathcal{L}^r_{mask} + \mathcal{L}^c_{mask} + \lambda_{id}\mathcal{L}_{id} + \lambda_{attn}\mathcal{L}_{attn}, \tag{10}$$

where $\lambda_{id} = 1$ and $\lambda_{attn} = 2.5$ is used for all experiments. For λ_{id}, 0.25 or 0.5 are also fine.

DreamBooth We employ the optimal settings for DreamBooth [22] training, which include prior preservation with a lambda value of 1.0 and a dataset of 200 images. Each batch comprises two images, consisting of one source and one style image. We set the learning rate to 1e-6 and train the model for 2000 steps. During this training, the CLIP text encoder and the UNet parameters are concurrently optimized. For inference, we use a denoising step of 50 with a guidance scale of 7.5.

Custom Diffusion For training the Custom Diffusion model [14], we use the best settings with prior preservation, a lambda value of 1.0, and a dataset of 200 images. The batch size is set to 2. With a learning rate of 5e-6, we train the model for 750 steps, optimizing the cross-attention layers of the Stable Diffusion model [21], as detailed in the original paper. The inference phase employs a denoising step of 50 and a guidance scale of 7.5.

Textual Inversion For the training of Textual Inversion [6], we adopt the optimal settings, including a batch size of 2, a learning rate of 5e-3, and a total of 5000 training steps. The inference process involves a denoising step of 50 with a guidance scale of 7.5.

General Object

Our method can also be applied to other general objects, where our composed prompt learning can be applied for robust multi-concept composition. We illustrate this with an example where the goal is to maintain the structure of the source object while adopting the texture from the reference image. We employ the same masked reconstruction objective $\mathcal{L}^s_{mask}$ for the source, and naive reconstruction objective without masking $\mathcal{L}^r$ for the reference.

For composed prompt learning, we employ structure loss [15] that maximizes structural similarity between the estimated image $\hat{x}^{(0)}$ and the source images using a pre-trained DINO ViT [3]. Specifically, the structure loss comprises two components: the self-similarity loss $\mathcal{L}_{ssim}$ [29] and the patch contrastive loss $\mathcal{L}_{contra}$ [17]. $\mathcal{L}_{ssim}$ utilizes a self similarity matrix derived from the multi-head self attention (MSA) layer of the pre-trained DINO. $\mathcal{L}_{contra}$ maximizes the patch-wise similarity between the keys of the source and the estimated image $\hat{x}^{(0)}$, with the keys extracted from the MSA layer of DINO. For the style similarity loss $\mathcal{L}_{style}$, we minimize the distance between DINO ViT [CLS] token embeddings of the reference and the estimated image $\hat{x}^{(0)}$. To sum up, our loss function for composed prompt learning is:

$$\lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{contra}\mathcal{L}_{contra} + \lambda_{style}\mathcal{L}_{style}, \tag{11}$$

where $\lambda_{ssim} = 0.1$, $\lambda_{ssim} = 0.2$, and $\lambda_{ssim} = 2$ is used for training.

We demonstrate the results for general objects in Fig. 11. Additionally, in Fig. 8, we provide a comparison with results produced without the use of composed prompt learning. These comparisons reveal that, in the absence of composed prompt learning, the outcomes tend to suffer from two main issues: either the structure of the source concept is inadequately preserved, or the style of the reference images is not effectively incorporated.

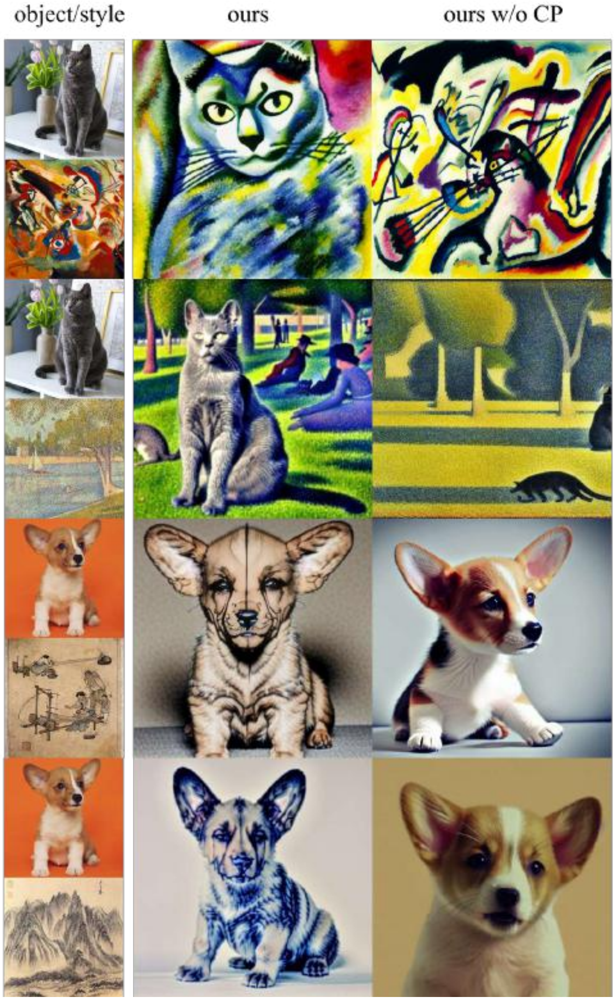


Figure 8: A comparison with results produced without the use of composed prompt learning for non-human images.

Ablation Study

We present the results of our ablation study in Table 3, which clearly highlight the significance of composed prompt learn-

Method	CSIM ↑	Style ↑	Aesthetic ↑
Ours	**0.566**	0.730	**6.218**
Ours w/ postprocessing	0.508	**0.737**	6.184
Ours w/o CP	0.429	0.717	6.159
Ours w/o AR & CP	0.429	0.726	6.178

Table 3: The results of the ablation study clearly highlights significance of composed prompt learning (CP) in enhancing the metrics. When CP is not included, there is a noticeable decline in CSIM and style score (measured by masked CLIP similarity).

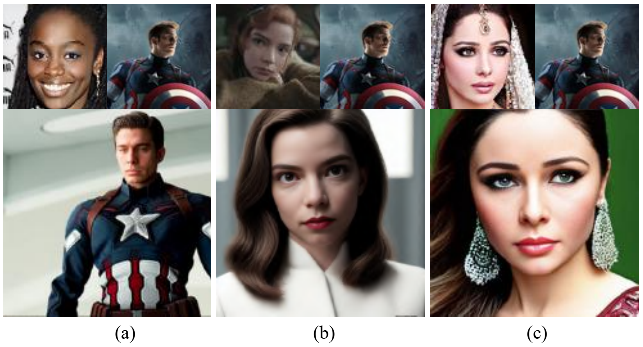


Figure 9: Results without Attention Refocusing (AR) loss. While AR loss does not appear to contribute to the metric improvement, the absence of AR often leads to *collapsed* samples as seen in (a) and (b). The generated samples predominantly reflect either the source or reference images, rather than a balanced combination of both. (c) illustrates that without AR, information spill is evident in the generated earrings, indicating that the source special tokens attend to non-facial regions.

ing (CP) in enhancing the metrics. When CP is not included, there is a noticeable decline in CSIM and style score (measured by masked CLIP similarity). Conversely, while the Attention Refocusing (AR) loss does not appear to contribute to the metric improvement, it is noteworthy that the absence of AR often leads to *collapsed* samples, where the generated samples predominantly reflect either the source or reference images, rather than a balanced combination of both. Illustrative examples of this are provided in Fig. 9, where Fig. 9 (a) showcases results that lean heavily towards the reference images, while Fig. 9 (b) exhibits only the source identity. Additionally, we observed instances of information spill when AR loss is not applied. Fig. 9 (c) illustrates that without AR, information spill is evident in the generated earrings, indicating that the source special tokens attend to non-facial regions. Finally, we note that the CSIM score exhibits a minor decline following post-processing. Although the post-processed results are generally visually appealing, the face restoration model possesses a level of freedom that can occasionally lead to a slight reduction in the similarity score. The results of samples before and after applying the post-processing are displayed in Fig. 10.

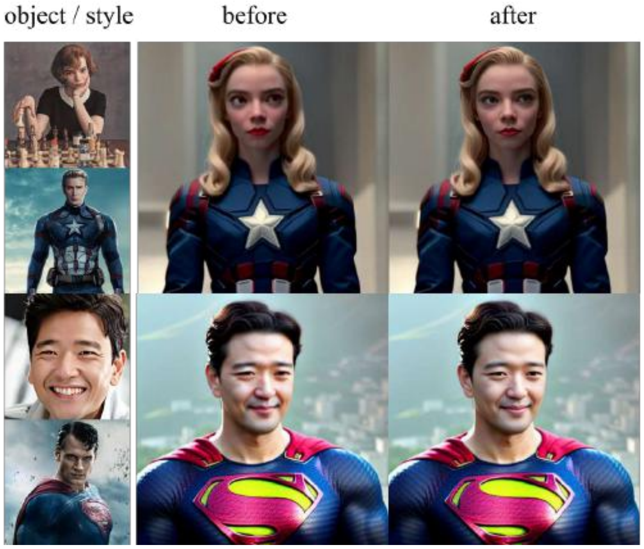


Figure 10: Generated results before and after post-processing.

Curated Results

We demonstrate more results from Fig. 12 to Fig. 17.

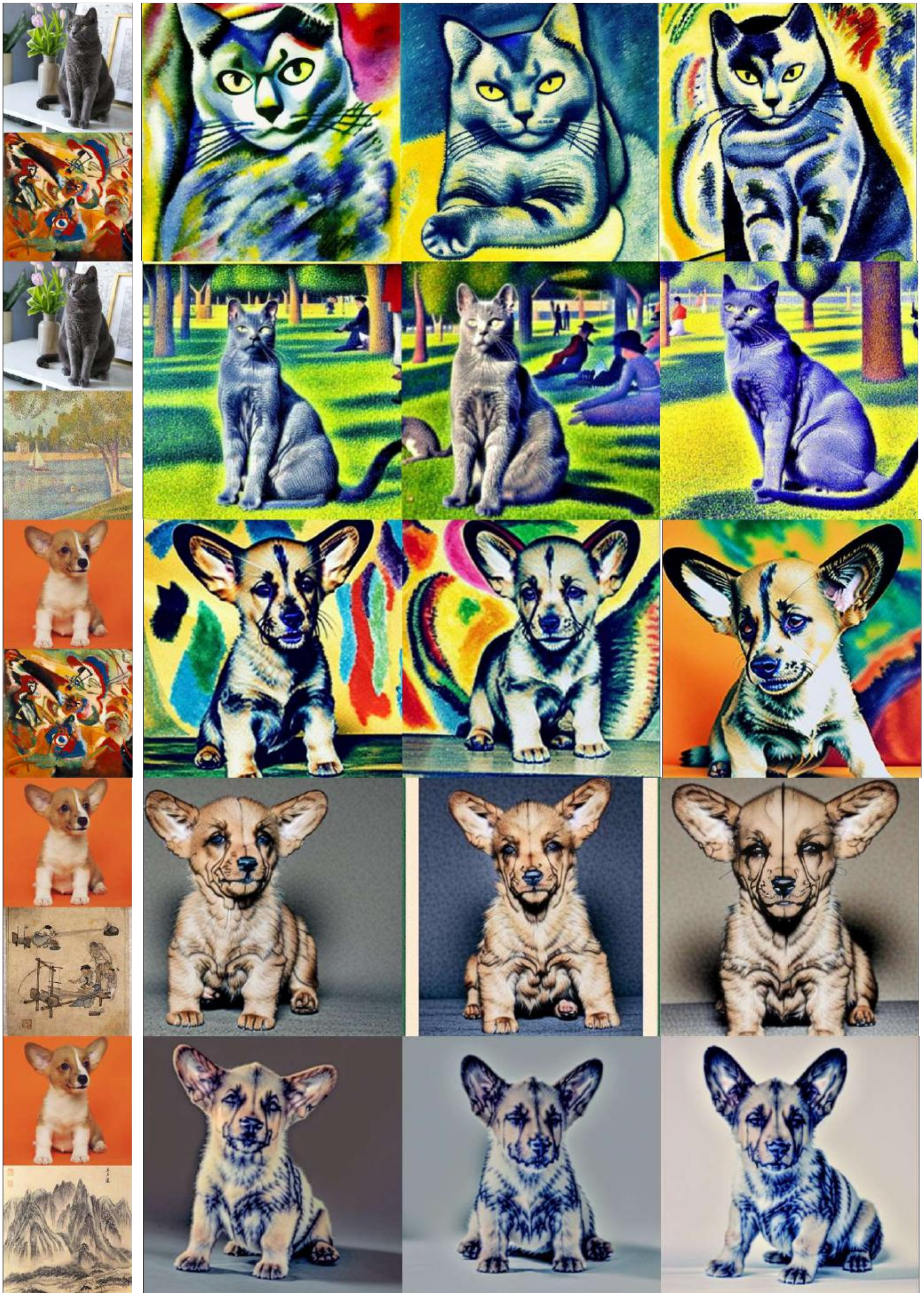


Figure 11: Results for composing the source content and the reference style in non-human images.

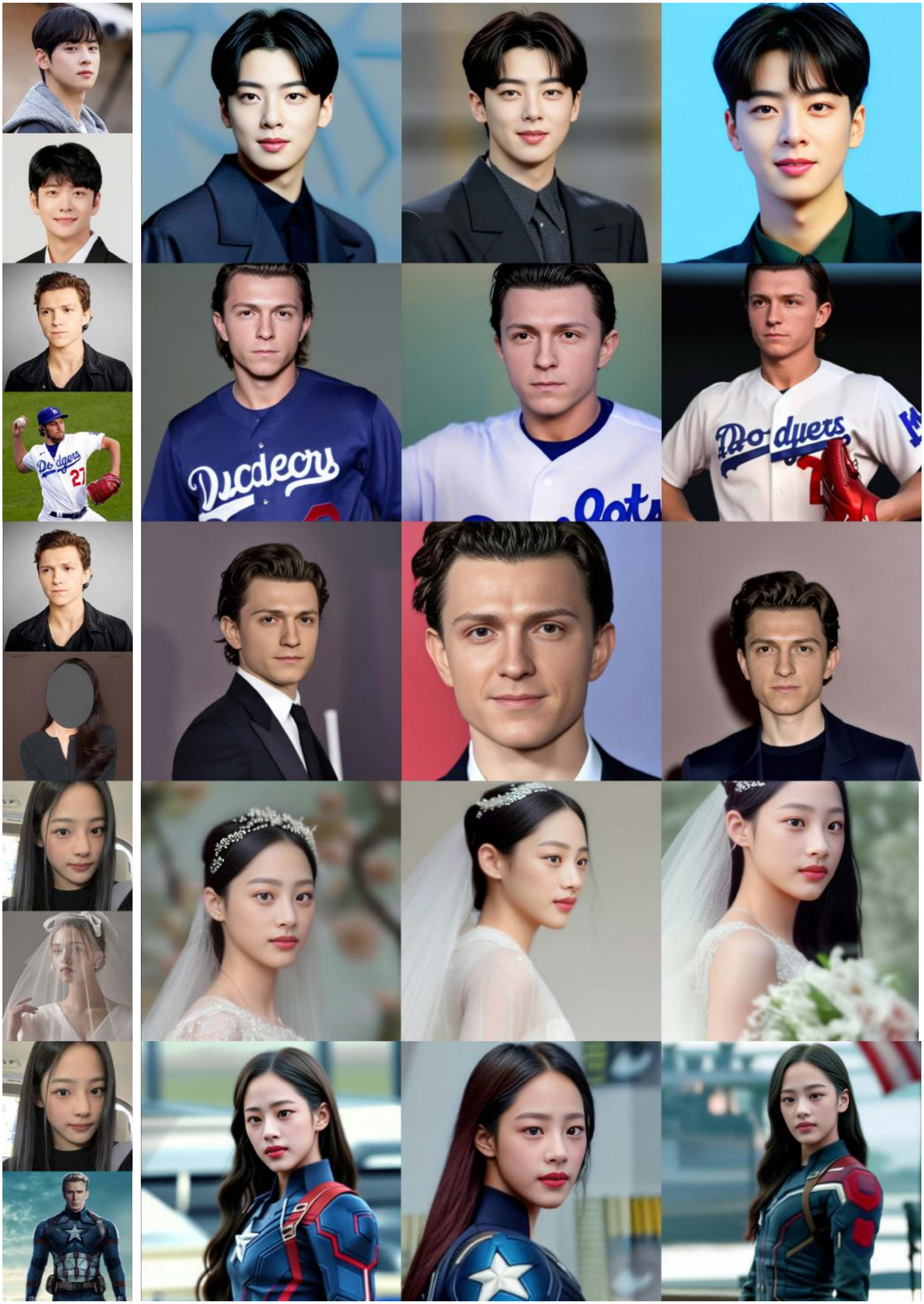


Figure 12: Curated results of MagiCapture.

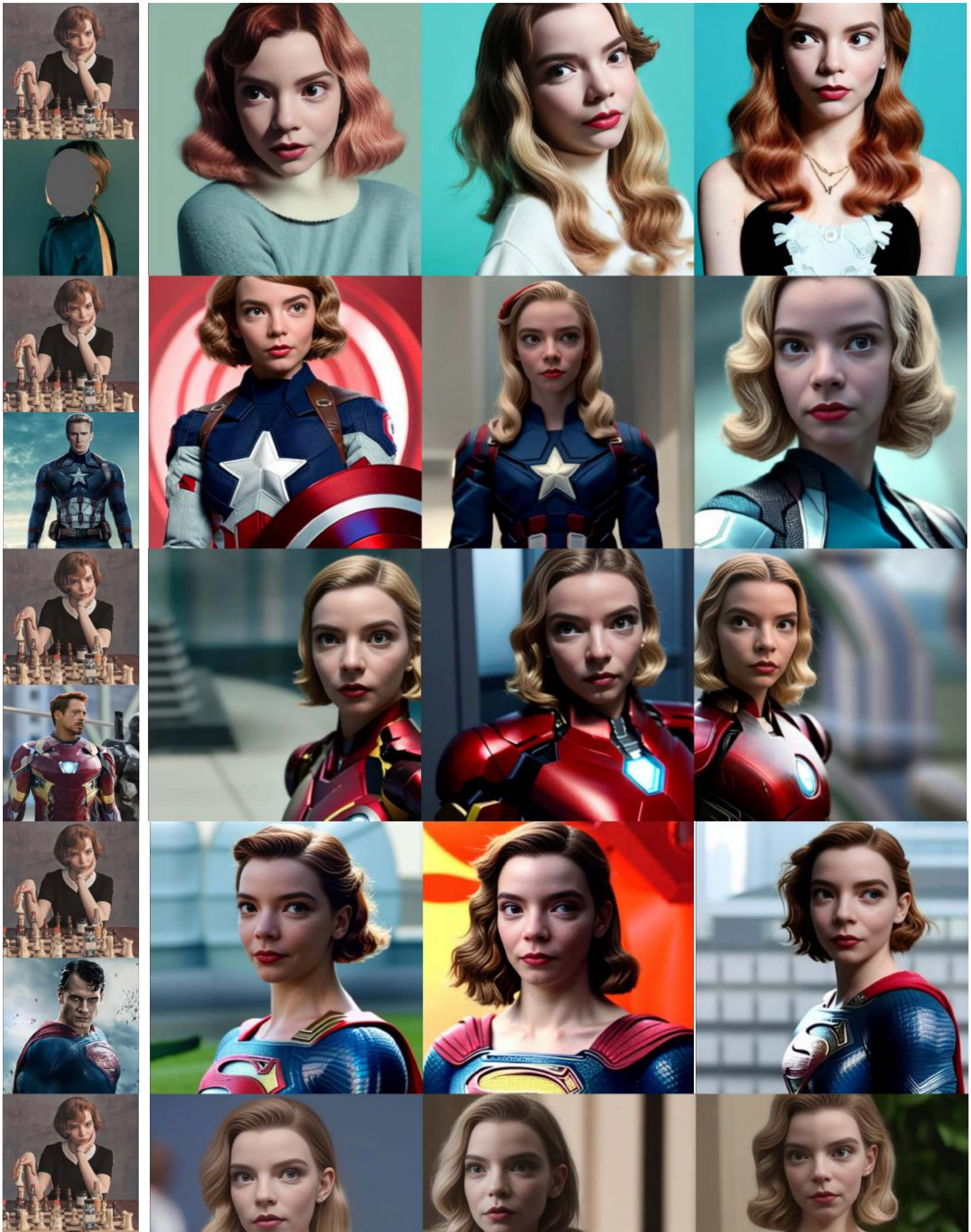


Figure 13: Curated results of MagiCapture.

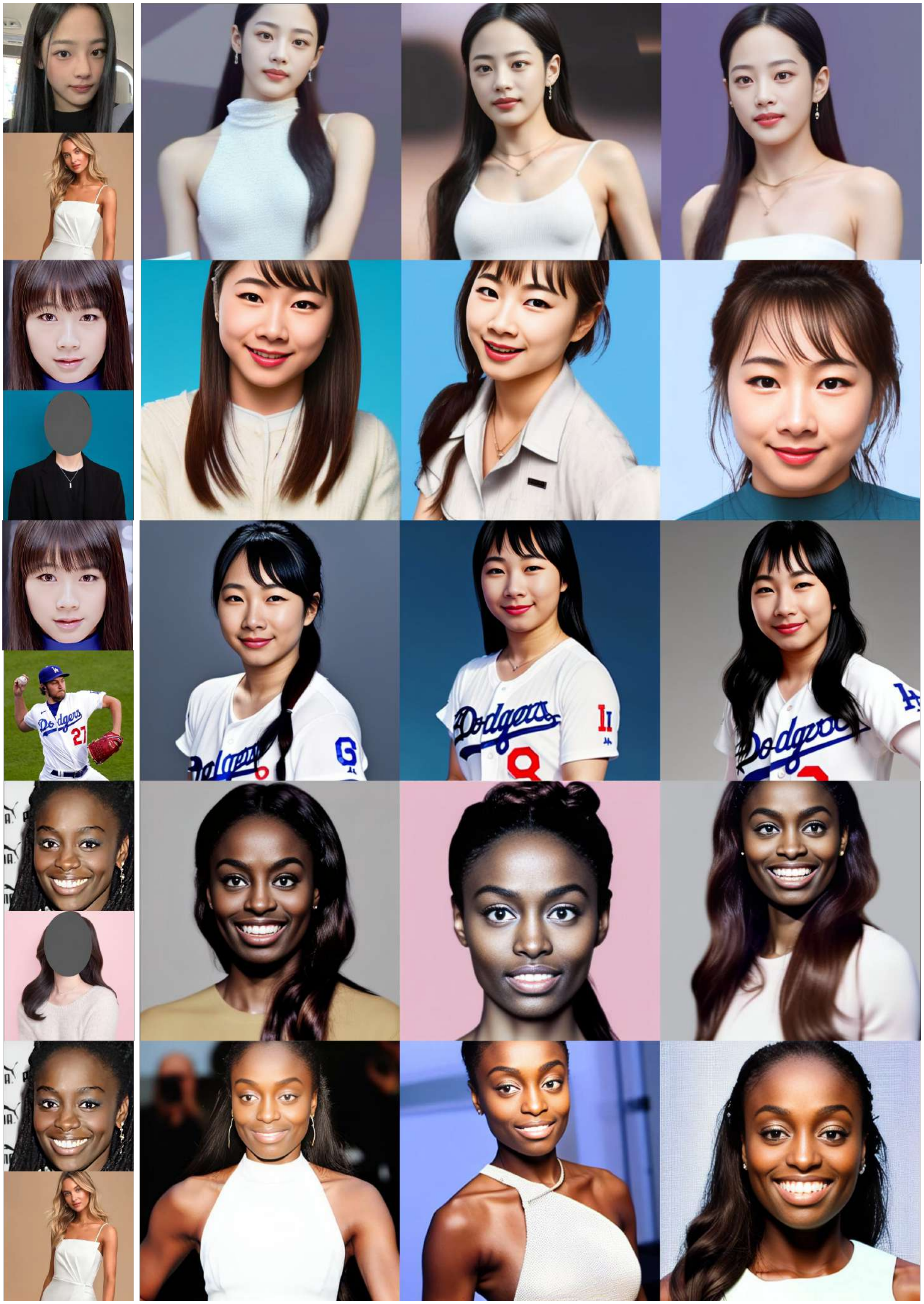


Figure 14: Curated results of MagiCapture.

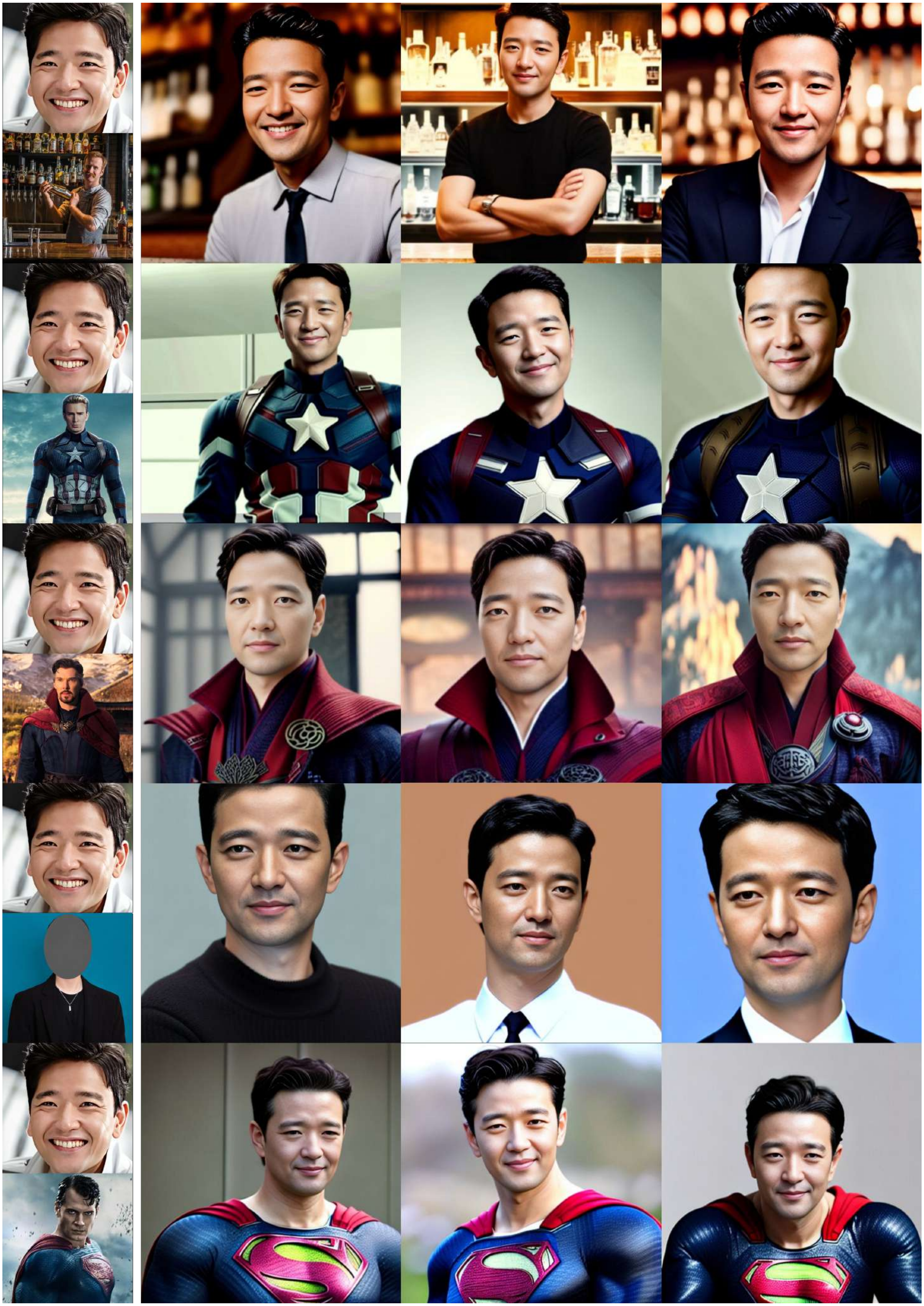


Figure 15: Curated results of MagiCapture.

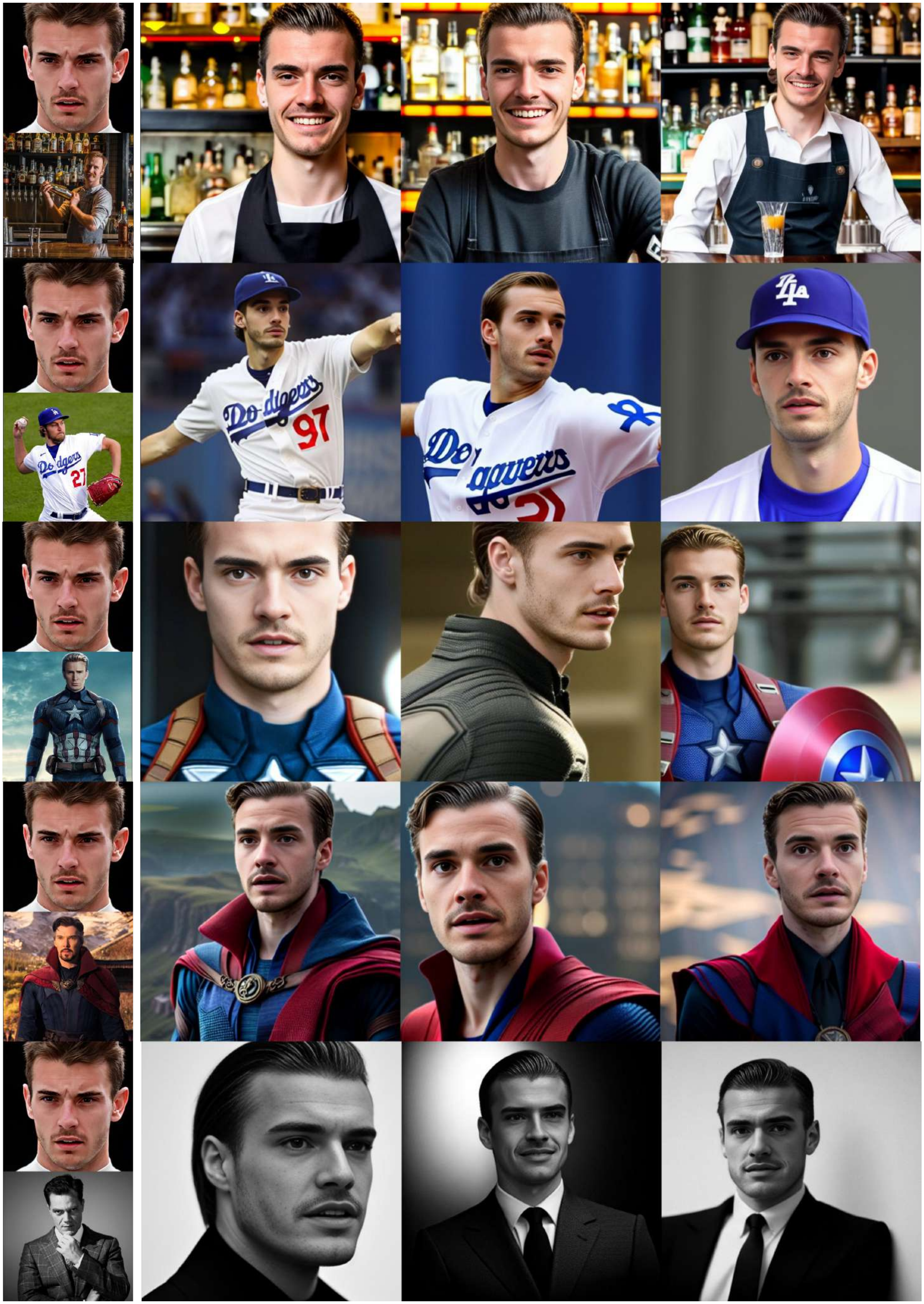


Figure 16: Curated results of MagiCapture.

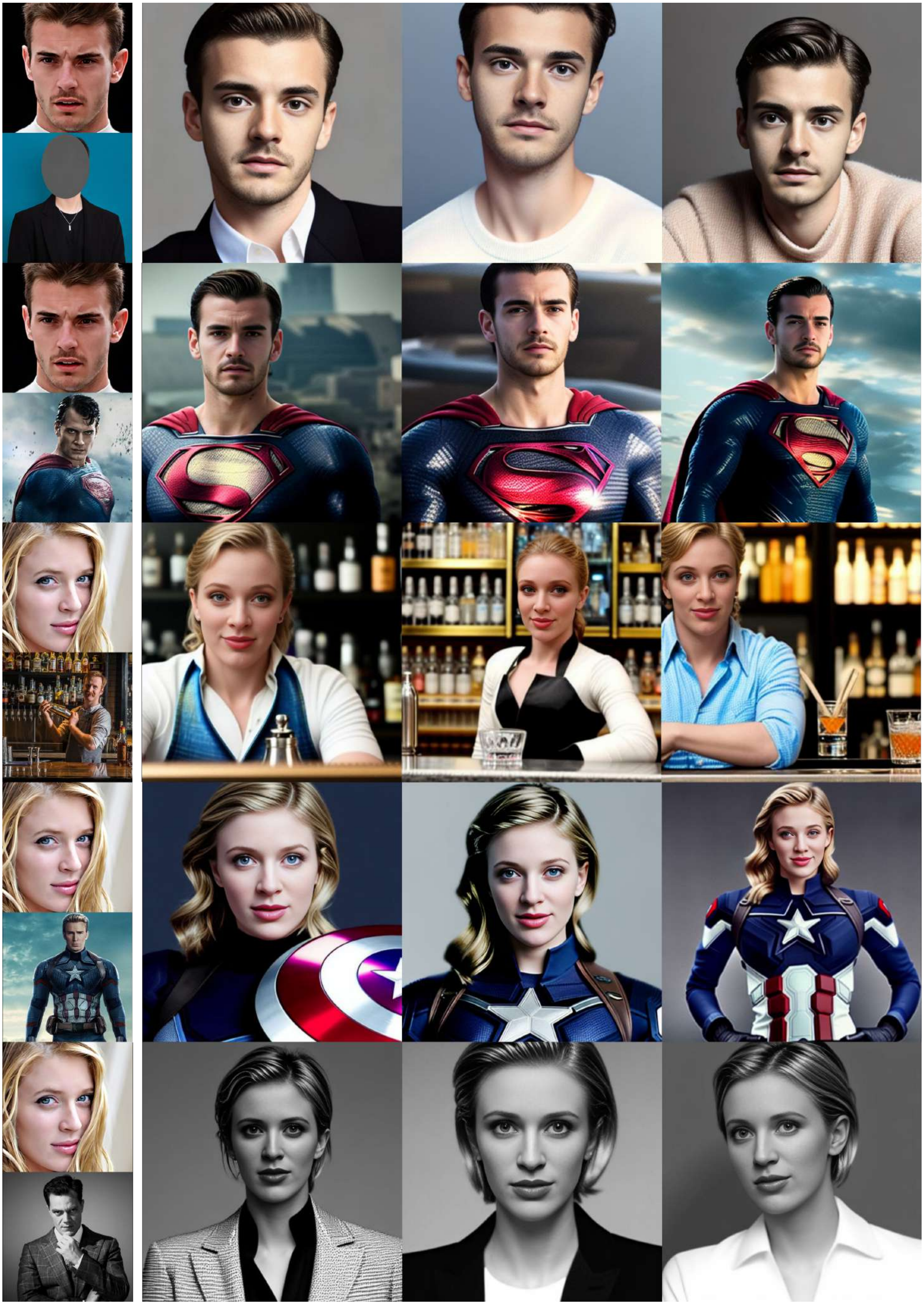


Figure 17: Curated results of MagiCapture.