

Say Cheese! Detail-Preserving Portrait Collection Generation via Natural Language Edits

Zelong Sun , Jiahui Wu , Ying Ba , Dong Jing , Zhiwu Lu *

Gaoling School of Artificial Intelligence
Renmin University of China, Beijing, China
zelongsun@ruc.edu.cn, luzhiwu@ruc.edu.cn

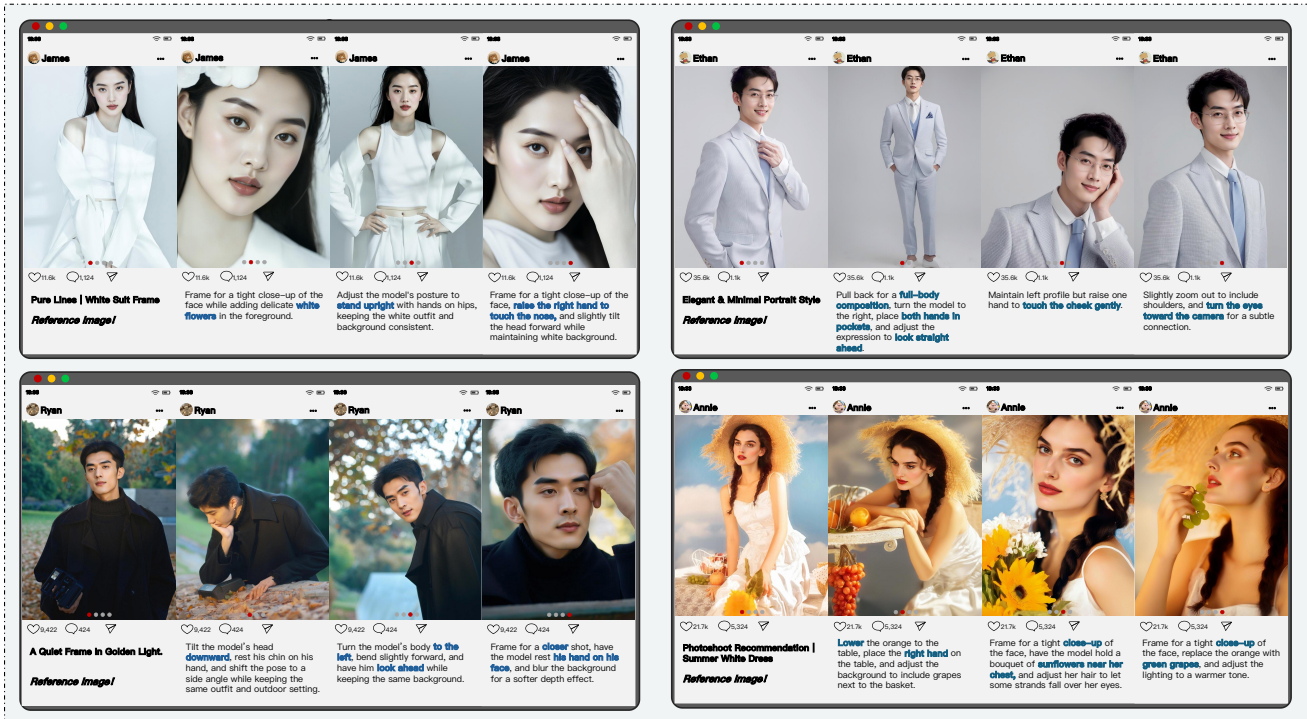


Figure 1. **Generated Portrait Collection Generation (PCG) Examples.** Each group shows a reference image (left) and three generated results by our method with corresponding modification texts (below the photos). PCG introduces two key challenges: (1) **Complex modifications** requiring simultaneous changes in pose, camera viewpoint, and spatial layout; (2) **Detail preservation** maintaining fine-grained appearance characteristics, such as makeup, clothing and accessories.

Abstract

As social media platforms proliferate, users increasingly demand intuitive ways to create diverse, high-quality portrait collections. In this work, we introduce Portrait Collection Generation (PCG), a novel task that generates coherent portrait collections by editing a reference portrait image through natural language instructions. This task poses two unique challenges to existing methods: (1) **complex multi-attribute modifications** such as pose, spatial layout, and camera viewpoint; and (2) **high-fidelity detail preservation** including identity, clothing, and accessories. To address these challenges, we propose **CHEESE**,

the first large-scale PCG dataset containing 24K portrait collections and 573K samples with high-quality modification text annotations, constructed through an Large Vision-Language Model-based pipeline with inversion-based verification. We further propose **SCheese**, a framework that combines text-guided generation with hierarchical identity and detail preservation. SCheese employs adaptive feature fusion mechanism to maintain identity consistency, and ConsistencyNet to inject fine-grained features for detail consistency. Comprehensive experiments validate the effectiveness of CHEESE in advancing PCG, with SCheese achieving state-of-the-art performance.

1. Introduction

In the digital era, portrait collections have become an essential way for people to document personal moments and express their identities, as shown in Figure 1. With the advancement of generative AI [8, 25, 26, 29], users increasingly demand intuitive ways to create diverse and high-quality portrait collections. To address this demand, we propose Portrait Collection Generation (PCG), a novel task that enables users to create portrait collections using a high-quality reference image and natural language instructions. Compared to existing image editing tasks, PCG introduces two key challenges, as illustrated in Figure 2: (1) **complex modification intents** spanning camera distance, pose, and viewpoint; and (2) **high-fidelity detail preservation** from the reference, including makeup, clothing, accessories, and fine identity cues, consistently across the collection.

Regarding the first challenge, ControlNet [37] and ControlNet++ [16] achieve fine-grained structure control by incorporating conditioning signals such as depth maps and Canny edges. However, these conditions often constrain spatial layout diversity, limiting the flexibility required for portrait photography. Other methods [3, 36], perform editing through text instructions by fine-tuning models on instruction-following datasets. However, their single-dimensional instructions often fail to meet users’ increasingly complex editing requirements. As shown in Figure 1 and 2, PCG frequently involve modifications in spatial layout, model pose, and viewpoint that are increasingly necessary to create diverse, expressive portrait collections.

Regarding the second challenge, existing fine-tuning methods such as DreamBooth [28] and LoRA [9] can improve consistency with reference images using a few reference samples. However, these methods require per-subject training, limiting their scalability in practical applications. Zero-shot alternatives like IP-Adapter [35] and InstantID [34] offer faster solutions but fall short in maintaining fine-grained details under complex modifications. Their fundamental limitation lies in their reliance on high-level semantic embeddings, which fail to preserve pixel-level details. PCG presents a more demanding challenge: simultaneously preserving both identity consistency and fine-grained detail consistency, illustrated in Figure 1 and 2.

To address these challenges, we first propose **CHEESE**, the first large-scale dataset specifically designed for PCG, emphasizing both complex multi-attribute modifications and high-fidelity detail preservation. We collect 24K high-resolution portrait collections from the web, encompassing diverse scenarios (e.g., indoor and outdoor), styles (e.g., ancient, ethnic, and modern), and variations (e.g., pose changes, camera angles/distance, and spatial layouts). We then pair images from the same portrait collection and employ a Large Vision-Language Model (LVLM) to obtain detailed, compositionally rich modification texts, resulting

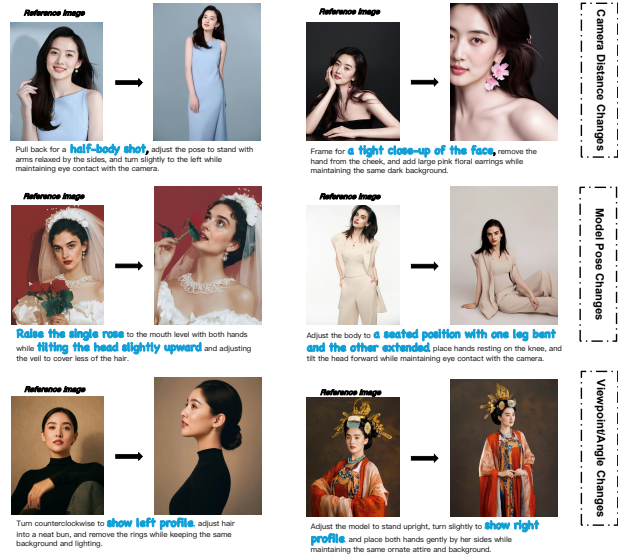


Figure 2. **Generated Examples of Several Editing Challenges in PCG:** camera distance, pose variation, and viewpoint transformation while preserving fine-grained details.

in 575K triplets. Note that modification texts can range from single-dimensional changes to complex combinations of multiple transformations. To ensure annotation quality, we introduce an inversion-based verification mechanism that provides quantitative scores for annotations and regenerates low-scoring ones. By sampling real-world, professional portrait collections, CHEESE ensures the **high-quality, diversity, creativity, content richness, and high-fidelity detail consistency** required for PCG.

We further propose **SCheese**, a framework that combines text-guided generation capabilities with identity- and detail-preserving mechanisms. For complex modification intents, we adopt a text-conditioned generation backbone that processes modification text to guide compositional transformations. For identity consistency, unlike IP-Adapter that only encodes the reference image, we fuse modification text and reference image features to directly obtain features that approximate the target image, supervised by an additional alignment loss. Such fused features provide more precise conditioning signals that align with the desired target characteristics, reducing ambiguity and improving instruction-following capability. During training, we apply teacher forcing by probabilistically replacing the fused features with target image features to provide direct supervision. For detail preservation, we propose ConsistencyNet, which extracts multi-scale fine-grained features from the reference image and injects them throughout the generation process via a Decoupled-Attention mechanism. Together, as shown in Figure 2, SCheese can balance complex modifications with fine-grained detail preservation.

Our main contributions are threefold: (1) We propose **CHEESE**, the first large-scale dataset for PCG featuring complex modification annotations and detail consistency;

(2) We design **SCheese**, a novel framework to balance complex modifications with fine-grained detail preservation; (3) Comprehensive experiments validate the effectiveness of CHEESE in advancing PCG, with SCheese achieving state-of-the-art performance in handling complex edits with identity and fine-grained details consistency.

2. Related Work

Text-Guided and Conditioned Image Editing. Text-guided image editing enables semantic modifications via natural language. Early diffusion-based approaches [14, 29] demonstrated text conditioning for generation. Recent works [3, 36] fine-tuned diffusion models on instruction-following datasets. However, these methods primarily handle simple edits and struggle with complex, multi-faceted instructions requiring simultaneous changes across pose, layout, and camera parameters [8, 26]. Beyond text-only control, methods like ControlNet [37] and T2I-Adapter [20] introduce spatial conditioning via sketches, depth maps, and segmentation masks. ControlNet++ [16] and Uni-ControlNet [39] further enhance spatial control integration. While these methods offer flexibility, they focus on structural conditioning or high-level semantics and lack capabilities for fine-grained identity preservation, which is critical for PCG task. Moreover, they typically operate on single-image edits without addressing consistency requirements for portrait collection generation.

Personalization in Diffusion Models. Image personalization approaches can be categorized into tuning-free methods [7, 12, 17, 24, 33, 33, 34, 38] and tuning-based methods [6, 9, 28]. DreamBooth [28] fine-tunes diffusion models on reference images to learn subject-specific representations, while LoRA [9] offers parameter-efficient fine-tuning via low-rank matrices. However, both require per-subject training and multiple references, making them computationally intensive. IP-Adapter [35] enables zero-shot image-conditioned generation via lightweight adapters, and InstantID [34] improves facial identity preservation using face recognition features. IPAdapter-Instruct [27] combines image prompts with instructions but relies on carefully crafted prompts. Alternative approaches [5, 10, 19] depend on manually curated datasets or are limited to simple attribute copying. These limitations become evident in PCG, where methods must simultaneously accommodate multi-attribute modifications while maintaining fine-grained detail preservation across collections.

3. CHEESE

In this section, we propose a large-scale dataset tailored for PCG, CHEESE $D = \{(I_r^{(n)}, T_m^{(n)}, I_t^{(n)})\}_{n=1}^N$, where I_r and I_t denotes the high-resolution reference and target image respectively, T_m denotes the complex modification text, and

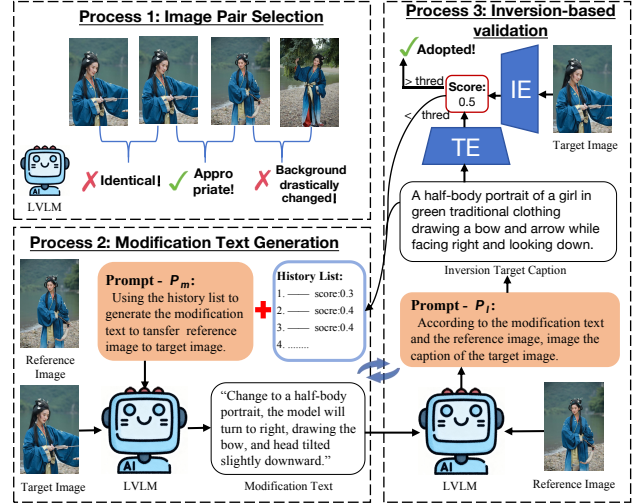


Figure 3. **The Dataset Construction Pipeline of CHEESE.** (1) We first use the LVLM filters out near-duplicate pairs and pairs with excessive background changes. (2) For each filtered pair, an LVLM generates a natural-language modification text describing the transformation. (3) We then use the LVLM to inversion the target caption and compute CLIP score with the target image. We then iteratively refine the annotation if the score falls below a threshold. IE , TE denote the image and text encoder of CLIP.

N denotes the triple number in D . The overview of our data construction pipeline in shown in Figure 3.

Image Pair Selection. We first collect portrait collections $A = \{A_k\}_{k=1}^K$ from the web, ensuring that each collection A_k contains multiple high-resolution images of a consistent subject and style. For each collection A_k , we enumerate all within-collection pairs (I_i, I_j) with $i \neq j$, and apply a LVLM to filter out near-duplicate pairs and those with excessive background or scene drift. This step removes trivial correspondences that would not support meaningful edits, while also discarding pairs where scene mismatch could confound identity preservation or detail fidelity. The result is a set of candidate pairs that balance visual diversity with subject and context coherence.

Modification Text Annotation. As shown in Figure 3, given a filtered pair $(I_r^{(n)}, I_t^{(n)})$, the annotation stage prompts an LVLM to generate a modification text $T_m^{(n)}$ describing the transformation from I_r to I_t . The prompt P_m design explicitly encourages the model to capture changes in camera parameters (such as depth of field and viewpoint), spatial/layout adjustments, and subject-level variations including pose, expression, and orientation. This design promotes compositional coverage and helps disentangle multi-faceted edits that often co-occur in PCG. Notably, given varying degrees of differences between image pair, T_m can range from single-dimensional changes to combinations of multiple transformations, aligning the diverse editing requirements in real-world.

Inversion-based Validation. To improve annotation ac-

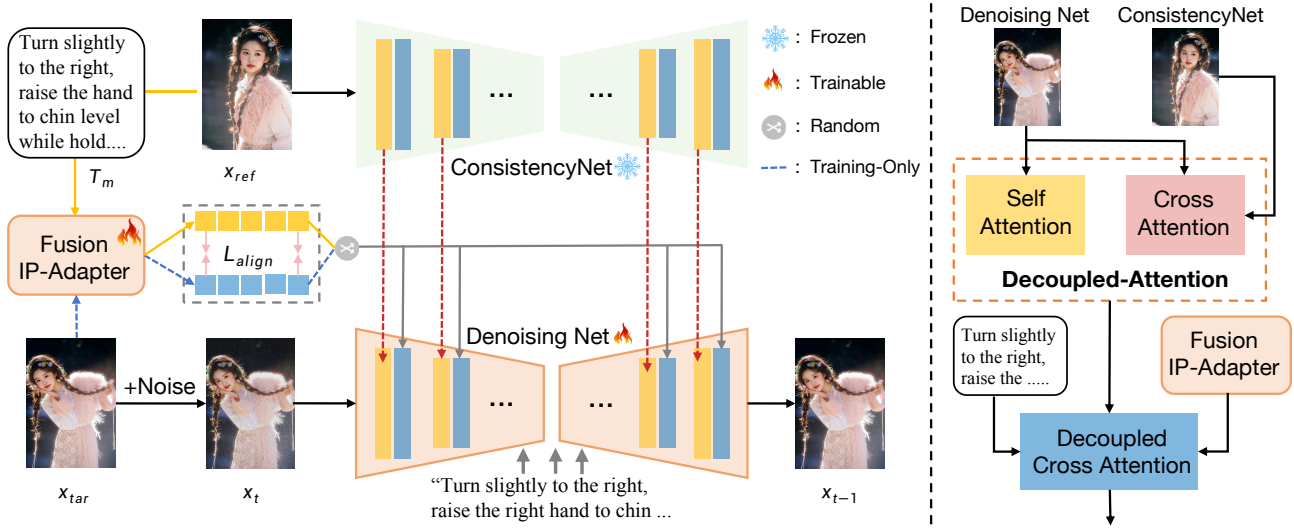


Figure 4. **Overview Architecture of SCheese:** (Left) Our model consists of (1) DenoisingNet which is a main UNet that processes target image, (2) Fusion IP-Adapter that fuses high-level semantics of reference image I_r and modification text T_m , and (3) ConsistencyNet that encodes low-level features of I_r . (Right) We propose a Decoupled-Attention mechanism, which consists of a self-attention module and a cross-attention module. The outputs of these two modules are averaged together and further fused with features from the text encoder and Fusion IP-Adapter through a decoupled cross-attention layer.

curacy, we design an inversion-based validation process, inspired by [13, 31]. Concretely, as shown in Figure 3, given $(I_r^{(n)}, T_m^{(n)})$, we ask the LVLm to generate an inversion target caption $\hat{c}$ that predicts the expected content of $I_t^{(n)}$. We then compute a CLIP-based similarity score $s = \cos(f_I(I_t^{(n)}), f_T(\hat{c}))$, where f_I and f_T denote the image and text encoders. We set the threshold τ a priori; if $s > \tau$, $T_m^{(n)}$ is accepted. Otherwise, we use the failed attempt’s $(T_m^{(n)}, s)$ as feedback to re-prompt the LVLm, re-generating a refined $T_m^{(n)}$. The process terminates either upon reaching the threshold or after M attempts. This iterative mechanism increases both the specificity and correctness, especially for complex, multi-attribute edits.

4. SCheese

4.1. Backgrounds on Diffusion Models

Denoising Diffusion Probabilistic Models [8] (DDPMs) are trained by maximizing the log-likelihood of the training data, given a data distribution $q(x_0)$. The training process involves a forward diffusion process that gradually adds Gaussian noise to the data over T timesteps:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) \quad (1)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I) \quad (2)$$

Here, x_t represents the noisy data at timestep t and α_t is a schedule parameter controlling the noise level at each

timestep. The core of DDPM training lies in learning a parameterized model p_θ to approximate the reverse diffusion process:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \quad (3)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (4)$$

This model learns to progressively remove noise from a given noisy sample x_t , recovering the original data x_0 .

4.2. Network Architecture

Overview. In this section, we present our method for designing diffusion models for PCG. Figure 4 provides the overview of our method. Given a triple (I_r, T_m, I_t) , our primary goal is to generate the target image I_t that satisfies the requirement of the modification text T_m and preserve the detail from the reference image I_r . To achieve that, we first adopt Stable Diffusion, a text-to-image generative model, as our backbone, and use T_m as the condition. We then propose two module to ensure the ip consistency and detail preserve: Fusion IP-Adapter and ConsistencyNet with decoupled attention.

Fusion IP-Adapter. To condition the high-level semantics of the reference image, we leverage an Image Prompt Adapter (IP-Adapter). Typically, the IP-Adapter consists of an image encoder f_{img} , a projection network that maps image features to the latent space of the diffusion model, and integration via cross-attention to inject image conditions into the denoising process. To obtain more accurate condi-

tions, inspired by Composed Image Retrieval (CIR), we additionally introduce a text encoder to extract semantic information from the modification text, and fuse the two modalities via a projection layer. Specifically, we extract image features $f_r = f_{\text{img}}(I_r)$ and text features $f_m = f_{\text{txt}}(T_m)$, then fuse them through a learnable projection network:

$$f_{\text{fused}} = \text{Proj}([f_r; f_m]), \quad (5)$$

where $[\cdot; \cdot]$ denotes concatenation and Proj is a query-based projection module. To optimize the fusion quality, we introduce a KL divergence loss between the fused features and target image features:

$$\mathcal{L}_{\text{align}} = \text{KL}(f_{\text{fused}} \| f_{\text{img}}(I_t)) \quad (6)$$

During training, we apply **teacher forcing** by probabilistically (with hyperparameter *pro*) replacing the fused features with the target image features. Since the fusion module produces features that approximate target features, teacher forcing provides two training modes: target features offer precise supervision signals for better model learning, while fused features simulate the inference scenario. This strategy enables the model to benefit from stronger supervision while maintaining generalization capability for inference, which further improves model performance.

ConsistencyNet. While we already condition the reference image using Fusion IP-Adapter, it falls short in preserving the fine-grained details when the reference image has complicated patterns or graphic prints. To tackle this issue, we propose to utilize an additional UNet to extract low-level features of the reference image. As show in Figure 4, we use another pretrained UNet encoder, ConsistencyNet, to obtain the intermediate representation of I_r , and adopt a decoupled-attention mechanism to inject these features into the generation process. Specifically, at each U-Net block, we add a parallel cross-attention layer alongside the self-attention module. The cross-attention layer receives queries from I_t and keys/values from I_r . We then average the self-attention and cross-attention outputs, and further fuse it with features from the text encoder and Fusion IP-Adapter through a decoupled cross-attention layer.

This decoupled design keeps self-attention focused on spatial dependencies within the generated image, while cross-attention handles explicit alignment between the generation state and the reference features. By maintaining separate feature spaces and combining them additively, the model can selectively attend to reference details without disrupting the generative flow. Notably, we adopt the in-painting model as ConsistencyNet and remove the mask-related layers to maximize the preservation of fine-grained details from the reference image. This mechanism of ConsistencyNet preserves intricate patterns that require pixel-level correspondence, while allowing the generation to adapt flexibly to modification instructions.

Method	VLM-based			LVLM-based	
	CLIP-I	DINO-I	CLIP-T	Qwen-DP	Qwen-PF
DreamBooth	0.642	0.636	0.375	0.305	0.443
DB LoRA	0.738	0.677	0.395	0.458	0.579
IP-Adapter	0.764	0.663	0.386	0.464	0.511
IP-Adapter+	0.794	0.699	0.376	0.659	0.549
EasyRef	0.783	0.687	0.358	0.647	0.545
BLIP-Diffusion	0.659	0.627	0.324	0.213	0.289
Emu2	0.849	0.821	0.379	0.767	0.352
Kolors	0.853	0.824	0.406	0.808	0.428
Kontext	0.857	0.791	0.413	0.792	0.679
SCheese (ours)	0.839	0.773	0.436	0.855	0.793

Table 1. **Quantitative Comparison Results.** SCheese achieves the highest scores on CLIP-T, DP, and PF, demonstrating superior capability in following modification instructions while preserving fine-grained details. While IP-Adapter (Kolors) and FLUX.1 Kontext achieve slightly higher CLIP-I and DINO-I scores, this likely reflects their “copy-pasting” tendency considering their lower PF score. **Bold** values indicate the best performance for each metric.

5. Experiments

5.1. Dataset

The CHEESE dataset comprises a large-scale collection of portrait albums with high-quality modification annotations. To avoid identity leakage, we perform collection-level splitting between training and testing. For the training set, we enumerate all valid within-collection pairs to expose the model to a wide range of modification patterns. The final training set contains $\sim 24\text{K}$ portrait albums, encompassing $\sim 40\text{K}$ images and $\sim 576\text{K}$ triplets. For the test set, we sample exactly one triplet per identity collection to maximize identity diversity and reduce redundancy, thereby providing a more stringent evaluation of generalization. The final test set consists of $\sim 2\text{K}$ triplets. All images maintain a minimum resolution of (832, 1216) pixels and are uniformly resized to (832, 1216) for consistent settings.

5.2. Evaluation Metrics

Following prior work [5, 28, 35], we first evaluate generated images using CLIP and DINO feature similarities. Specifically, we use ViT-G/14 CLIP from OpenCLIP [11] and DINOv2-Small [21] to calculate similarities between I_r and generated images I_g , denoted as CLIP-I and DINO-I, to measure detail preservation. We also report CLIP-T, the similarity between I_g and the inversion target caption (produced by Process 3 in Figure 3), to evaluate how well the modification text is followed.

However, these metrics primarily capture global semantic similarity, and since I_r and I_t already share high similarity, a model that simply copies I_r can achieve high scores, leading to misleading evaluations [4]. To address this limitation, we propose two LVLM-based evaluation metrics us-



Figure 5. **Qualitative Comparisons.** Blue markers highlight successful detail preservation and instruction compliance; red boxes indicate the “copy-pasting” tendency; orange boxes denote blurriness or detail loss; purple boxes indicate limb abnormalities. SCheese successfully maintains reference details and image realism while executing complex modifications.

ing Qwen-2.5-VL 72B [1] to separately assess detail preservation (DP) and prompt-following (PF) capabilities. In our prompts, we also incorporate debiasing mechanisms that penalize “copy-pasting” behavior as in [4]. This evaluation protocol emulates human evaluation via LVLMs. The full set of evaluation prompts are provided in **Appendix**.

5.3. Implementation Details

For dataset curation, we adopt Qwen2.5-VL 72B [1] as our LVLm, and ViT-G/14 CLIP from OpenCLIP [11] as the text and image encoder. The hyperparameter τ and M is set to 0.45 and 5.

For our method, SCheese, we use SDXL [23] as our Denoising Net, SDXL inpainting model for ConsistencyNet and IP-Adapter [35] plus for our Fusion IP-Adapter, and initied them with Kolors [32] weight. *pro* is setting to 0.35. We train all models on 8 NVIDIA H800 80GB GPUs with an effective batch size of 64 for 50k iterations, using AdamW optimizer [18] with a learning rate of 1e-5.

5.4. Baselines

Following [4, 22], we compare our model with two classes of baselines: inference-stage fine-tuning models and zero-shot models. For inference-stage fine-tuning models, we compare against DreamBooth [28] and its LoRA [9] version (denoted as DB LoRA). For zero-shot models, we compare with BLIP-Diffusion [15], Emu2 [30], IP-Adapter [35], IP-Adapter+ [35], IP-Adapter with Kolors initialization (denoted as Kolors) [32], EasyRef [40], and FLUX.1 Kontext [2].

5.5. Comparison with Baselines

Quantitative Results. Table 1 presents quantitative evaluation results comparing SCheese against existing baselines on the CHEESE test set. From the results, we observe that: (1) Methods requiring per-subject fine-tuning, including DreamBooth, exhibit limited performance across evaluation metrics, whereas zero-shot approaches such as IP-Adapter demonstrate competitive results without requir-

Metric	Target	Emu2	Kolors	Kontext	SCheese
Human-DP	0.936	0.158	0.410	0.653	0.778
Human-PF	0.930	0.138	0.397	0.670	0.803
Human-CC	0.915	0.115	0.293	0.467	0.688

Table 2. **User Study.** “Target” denotes the target image, which serves as an upper bound. Our user study results mostly align with our LVLM-based evaluation. **Bold** values indicate the best performance of models for each metric.

ing test-time adaptation. (2) Our method SCheese attains the best performance on Detail Preservation, Prompt Following, and CLIP-T, indicating its effectiveness in executing complex editing instructions while maintaining high-fidelity detail preservation. (3) Certain baselines, notably FLUX.1 Kontext and Emu2, achieve marginally higher reference similarity scores (CLIP-I and DINO-I) but exhibit substantially lower prompt following performance. This performance trade-off suggests these methods prioritize reference replication over instruction adherence, a finding corroborated by our qualitative analysis (see Figure 5). Collectively, these results demonstrate that SCheese achieves an optimal trade-off across detail preservation, instruction compliance, and semantic alignment, validating its suitability for PCG task.

Qualitative Results. Figure 5 presents our qualitative comparison results, demonstrating that our model outperforms all baselines in detail consistency while exhibiting excellent prompt alignment in the outputs. (1) As shown in the figure, DreamBooth [28] frequently fails to preserve consistent visual elements from the reference image, such as clothing and headwear. This performance gap likely results from its design assumption of multiple reference samples. (2) IP-Adapter [35] and IP-Adapter+ [35] has improving subject-level consistency, but their reliance on high-level embeddings prevents them from maintaining fine-grained details, as evidenced by the cake in the third row. (3) Emu2 [30] demonstrates strong semantic understanding capabilities but lacks dedicated mechanisms for subject-specific personalization. (4) Additionally, Emu2 shows limited comprehension of complex editing instructions, frequently generating images that closely mirror the reference without incorporating the requested modifications. This “copy-paste” tendency [4] is similarly observed in Kolors IP-Adapter and FLUX.1 Kontext, as highlighted by the red boxes in Figure 5. This tendency leads to high reference similarity, resulting in elevated CLIP-I and DINO-I scores, while yielding lower CLIP-T and PF scores due to their failure to follow the modification text requirements. (5) In contrast, our approach effectively preserves the subject’s core identity while enabling diverse and contextually appropriate transformations. As indicated by the blue boxes in Figure 5, our method successfully maintains details, such as the “cake

Reference Image	Target Image	w/o validate	ViT-B/32	ViT-G/14
		Replace the golden staff with a black sword held diagonally across the body while adjusting the pose to lean slightly forward.	Replace the golden staff with a black sword held diagonally across the body, adjust the pose to lean slightly forward, and turn the head to the left.	Replace the golden staff with a black sword held diagonally across the body, adjust the pose to lean slightly forward, and turn the head to the left while maintaining eye contact with the camera.
		Turn the model counterclockwise to show her left profile while keeping the same seated position and basket placement.	Turn counterclockwise to show the left profile, and move closer to the wooden wall while maintaining the seated position.	Turn counterclockwise to show the left profile, place the basket on the lap, and slightly move the red banner into the frame.
		Turn counterclockwise to show left profile while maintaining eye contact with the camera.	Turn counterclockwise to show the left profile while keeping the gaze directed toward the camera.	Turn counterclockwise to show the left profile while keeping the gaze directed toward the camera and changing the top to a sleeveless design.
		Extend the right leg outward, adjust the body to lean slightly forward, turn the head downward, and add a white door with a gold handle.	Extend the right leg outward, adjust the body to lean slightly forward, turn the head downward, and add a white door with a gold handle.	Extend the right leg outward, adjust the body to lean slightly forward, turn the head downward, and add a white door with a gold handle to the right side of the frame while keeping the left leg bent and resting on the edge of the furniture.

Figure 6. **Ablation Study for the Inversion-based Validation process.** Comparison of modification text quality across three annotation strategies. Color annotations (blue/red) highlight improvements over the baseline.

and white dress” in the third row, and even the “pearl earrings” in the last row, while also following the modification text requirements, including changes in model orientation, hand poses, and camera distance variations.

User Study. To evaluate the fidelity and prompt consistency of our generated images, we conducted a user study on a random subset of CHEESE test cases, selecting 50 samples. Users independently scored each image from 0 to 4 based on three criteria: DP, PF, and Collection Coherence (CC). CC measures users’ willingness to include the generated image in the same portrait collection as the reference image. We also include target images in the evaluation to show the upper bound of each metric. The average scores are presented in Table 2. We observe the following: (1) Target images achieve near-perfect PF scores, which validates that our annotation method produces comprehensive and effective modification texts. (2) SCheese achieves highest PF scores among models, demonstrating its capability to execute complex modification requirements. Meanwhile, thanks to Fusion IP-Adapter and ConsistencyNet, SCheese also achieves the highest DP score, indicating its effectiveness in preserving reference details. (3) According to CC scores, users show strong willingness to use SCheese-generated images to form portrait collections, demonstrating the practical potential of our method.

6. Ablation Study

Effect of Inversion-based Validation. In Figure 6, we investigate the impact of the Inversion-based Validation process during dataset curation. We compare three annotation strategies: (1) no validation process, (2) validation using CLIP with ViT-B/32, and (3) validation using CLIP with ViT-G/14. Our analysis reveals two key findings: (1) Com-

Method	VLM-based			LVLM-based	
	CLIP-I	DINO-I	CLIP-T	Qwen-DP	Qwen-PF
IP-Adapter (ZS)	0.853	0.824	0.406	0.808	0.428
+ SFT	0.822	0.728	0.421	0.783	0.683
+ ConNet	0.826	0.732	0.418	0.832	0.673
+ Fusion IP-A	0.828	0.732	0.416	0.828	0.723
+ Align Loss	0.837	0.753	0.426	0.836	0.777
+ Teacher (Full)	0.839	0.773	0.436	0.855	0.793

Table 3. **Ablation Study Results for Proposed Components.** **Bold** values indicate the best performance for each metric. Each component contributes to performance: +SFT improves instruction following, +ConNet enhances detail preservation, +Fusion IP-A and +Align Loss further refine capabilities, and +Teacher achieves optimal performance across all metrics.

pared to the baseline without validation, both CLIP-based validation approaches guide the LVLM to identify more nuanced differences between images, resulting in more comprehensive and detailed modification texts, as highlighted by the blue and red annotations in the figure. This demonstrates the effectiveness of the inversion-based validation mechanism in improving annotation quality. (2) When comparing the two CLIP variants, ViT-G/14 produces more refined and accurate modification texts than ViT-B/32, as indicated by the red annotations. This performance gap likely stems from ViT-G/14’s superior image-text understanding capabilities, which provide more accurate supervision signals for the validation process, enabling better discrimination between high-quality and low-quality annotations.

Effects of Components. To validate the contribution of each component in SCheese, we conduct incremental ablation experiments. Quantitative results are reported in Table 3, and qualitative results are shown in Figure 7. From Table 3, we observe the following: (1) Compared to the zero-shot baseline, supervised fine-tuning on the CHEESE dataset (+SFT) substantially improves the model’s ability to follow modification instructions, as reflected in CLIP-T and PF scores, while maintaining image quality. (2) Incorporating ConsistencyNet (+ConNet) leads to significant improvements in detail preservation (DP). However, we notice PF slightly decreases, likely because the more influence from the reference image. (3) Adding Fusion IP-Adapter (+Fusion) further enhances instruction-following capability, as the additional modification text input provides more guidance for the editing process. (4) Introducing the alignment loss ($\mathcal{L}_{\text{align}}$) encourages fused features to align more closely with target image, improving fusion quality and resulting in overall performance gains. (5) Finally, incorporating teacher forcing (+Teacher) provides more specific supervision signals for both the fusion module and denoising network, leading to further improvements.

Figure 7 demonstrates consistent findings. We highlight different modification intents in the modification texts us-



Figure 7. **Qualitative Ablation Results for Components.** Color annotations link modification intents (in text) to corresponding regions (in images).

ing distinct colors and annotate corresponding regions in the generated images with matching colored boxes. The qualitative analysis reveals: (1) +SFT enables the model to handle complex, multi-faceted instructions that were previously challenging, such as “seated position” and “hand on the lap”. (2) +ConNet allows the model to inherit fine-grained details from the reference image, such as headwear, and clothing. (3) With Fusion IP-Adapter, the model simultaneously satisfies multiple complex modification requirements (e.g., striped pillows and books) while maintaining high detail preservation with the reference.

7. Conclusion

In this work, we focus on the Portrait Collection Generation (PCG) task, which presents two fundamental challenges: complex multi-attribute modifications and high-fidelity detail preservation. To tackle these challenges, we propose SCheese, the first large-scale dataset for PCG containing 576K triplets with high-quality modification annotations generated through an LVLM-based pipeline with inversion-based verification, and SCheese, a framework that combines text-guided generation with hierarchical identity and detail preservation mechanisms. Experimental results demonstrate that SCheese outperforms existing methods on detail preservation and prompt following. We believe that our work establishes a solid foundation for advancing research in PCG and personalized image editing.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (62376274, 62437002).

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 6
- [3] 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, 3
- [4] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025. 5, 6, 7
- [5] Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023. 3, 5
- [6] 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. 3
- [7] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 3
- [8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 4
- [9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 3, 6
- [10] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multi-task learners. 2024. 3
- [11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, et al. Openclip. 2021. 5, 6
- [12] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, 37:76177–76209, 2024. 3
- [13] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 4
- [14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 3
- [15] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 6
- [16] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback: Project page: liming-ai. github. io/controlnet.plus.plus. In *European Conference on Computer Vision*, pages 129–147. Springer, 2024. 2, 3
- [17] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 3
- [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [19] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3
- [20] 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. 3
- [21] 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. 5
- [22] Yang 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. 6
- [23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6
- [24] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF con-*

- ference on computer vision and pattern recognition*, pages 8693–8702, 2024. 3
- [25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3
- [27] Ciara Rowles, Shimon Vainer, Dante De Nigris, Slava Elizarov, Konstantin Kutsy, and Simon Donné. Ipadapter-instruct: Resolving ambiguity in image-based conditioning using instruct prompts. In *European Conference on Computer Vision*, pages 54–70. Springer, 2024. 3
- [28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3, 5, 6, 7
- [29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 3
- [30] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyang Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 6, 7
- [31] Zelong Sun, Dong Jing, and Zhiwu Lu. Cotmr: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2502.20826*, 2025. 4
- [32] Kolos Team. Kolos: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 6
- [33] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 3
- [34] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3
- [35] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapt: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 5, 6, 7
- [36] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 2, 3
- [37] 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, 3
- [38] 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. 3
- [39] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023. 3
- [40] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *Forty-second International Conference on Machine Learning*, 2024. 6