Training-free Content Injection using h-space in Diffusion models

Supplementary Material

```
Algorithm 2: InjectFusion
  Input: $x_T$ (inverted latent variable from original image $I^{original}$),$\{h_t^{content}\}_{t=t_{edit}}^{T}$(obtained
         from content image $I^{content}$), $\epsilon_\theta$ (pretrained model), $m$ (feature map mask), $f$ (Slerp), $\omega$
         (calibration parameter)
  Output: $\tilde{x}_0$ (transferred image)
  $\tilde{x}_t \leftarrow x_T$ for $t = T, ..., 1$ do
      if $t \geq t_{edit}$ then
          // step1:  Content injection
          Extract feature map $h_t$ from $\epsilon_\theta(\tilde{x}_t)$;
          $\tilde{h}_t \leftarrow f((m \otimes h_t), (m \otimes h_t^{content}), \gamma), \omega$
                  $\oplus(1 - m) \otimes h_t$
          // step2:  Latent calibration
          $\tilde{\epsilon} \leftarrow \epsilon_\theta(\tilde{x}_t|\tilde{h}_t)$, $\epsilon \leftarrow \epsilon_\theta(\tilde{x}_t)$
          $\mu_{\mathbf{P}_t(\tilde{\epsilon})}, \sigma_{\mathbf{P}_t(\tilde{\epsilon})} \leftarrow \mathbf{P}_t(\tilde{\epsilon})$
          $\mu_{\mathbf{P}_t(\epsilon)}, \sigma_{\mathbf{P}_t(\epsilon)} \leftarrow \mathbf{P}_t(\epsilon)$
          $\mathbf{P}'_t = \mu_{\mathbf{P}_t(\tilde{\epsilon})} + (\mathbf{P}_t(\tilde{\epsilon}) - \mu_{\mathbf{P}_t(\tilde{\epsilon})}) * \sigma_{\mathbf{P}_t(\epsilon)}$
          $d\mathbf{P}_t = \mathbf{P}'_t - \mathbf{P}_t(\epsilon)$
          $d\epsilon = \tilde{\epsilon} - \epsilon$
          $dx = \sqrt{\alpha_t} * d\mathbf{P}_t + \omega * \sqrt{(1 - \alpha_t)} * d\epsilon$
          $\tilde{x}_t{}' = \tilde{x}_t + dx$
          $\tilde{\epsilon} = \epsilon \leftarrow \epsilon_\theta(\tilde{x}_t{}')$
      else
          $\tilde{\epsilon} = \epsilon \leftarrow \epsilon_\theta(\tilde{x}_t)$,
      $\tilde{x}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}(\frac{\tilde{x}_t - \sqrt{1-\alpha_t}\tilde{\epsilon}}{\sqrt{\alpha_t}}) + \sqrt{1 - \alpha_{t-1}}\epsilon$
```

A. Implementation details

To perform the reverse process for figures, we use 1000 steps, while for tables and plots, we use 50 steps. During inference, we injecte $\boldsymbol{h}_t$ sparsely only at the timesteps where the content injection applied within the 50 inference steps. For the remaining timesteps, we use the original DDIM sampling. This approach enables us to achieve the same amount of content injection across different inference steps.

For local mixing, we spatially apply Slerp on $\boldsymbol{h}_t$, which has a dimension of $8 \times 8 \times 256$, as demonstrated in Figure S1. In face swapping, we use a portion of $\boldsymbol{h}_t$ that corresponds to the face area for Slerp. In § 3, we use the editing interval [T=1000, t_{edit}=400], and do not use quality boosting to eliminate stochasticity for comparison purposes, i.e., $t_{\text{boost}} = 0$.

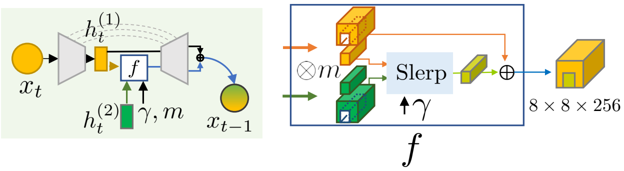


Figure S1. **Illustration of local mixing** Mask m determines the area of feature map. Slerp of masked $\boldsymbol{h}_t$ enables content injection into designated space.

B. Varying the strength of content injection

Figure S2 illustrates the results of content injection with different values of Slerp ratio γ. As observed in Figure 7b, there is a positive correlation between γ and the amount of content change. However, increasing $\gamma > 0.6$ barely leads to any content change but degrades the quality of images with distortions and artifacts. As the recursive injection of content by γ exponentially decreases the original $\boldsymbol{h}_t$ component along the reverse process, according to Eq. (13), we expect linear change of content in the image by linearly controlling α that specifies $\gamma = \alpha^{1/T}$.

C. Effect of latent calibration

In this section, we present an analysis of the parameter ω which specifies the strength of the original element. Figure S3 displays the resulting images with sweeping ω. As ω increases, the style elements become more prominent. We note that latent calibration with $\omega = 0$ is not rigorously defined and we report the results without latent calibration when $\omega = 0$. In Figure S4, we observe a trade-off between Gram loss and ID similarity, as well as FID, depending on the value of ω. However, despite this trade-off, increasing ω results in more effective conservation of the original image.

Because latent calibration also can control the strength of feature-injected results, we can utilize latent calibration for other feature-injecting methods, e.g., Plug-and-Play [71] and MasaCtrl [6]. Figure S5 shows that increasing ω increases the strength of editing.

D. More results and comparison

D.1. More qualitative results

We provide more qualitative results of CelebA-HQ, AFHQ, METFACES, LSUN-church, and LSUN-bedroom in Figure S18-S24 (located at the end for compact arrangement). We also provide a result of ImageNet in Figure S12a.

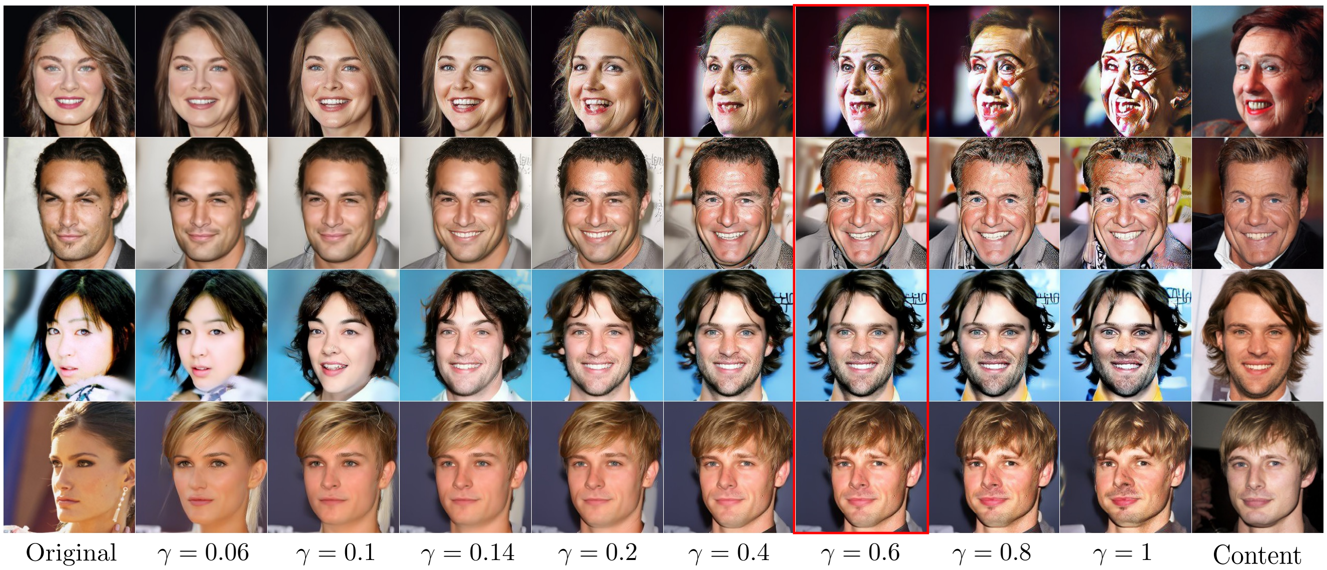


Figure S2. γ controls how much content will be injected. We do not use other techniques such as quality boosting for comparison.

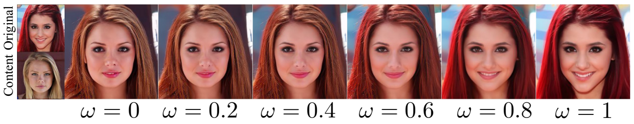


Figure S3. **Effect of increasing** ω. Increasing ω reflects style elements stronger and $\omega = 0$ shows the result without latent calibration.

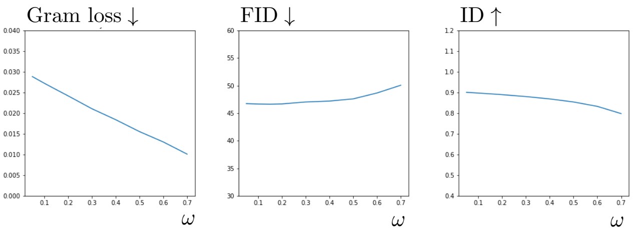


Figure S4. **Quantitative results of latent calibration with varying** ω**.** Latent calibration ensures that the resulting image remains close to the original image, minimizing content injection loss and preserving image quality.

	Method	Preference (%)
Content injection	Swapping Autoencoder [56]	40.11
	Ours	**59.89**
Local content injection	StyleMapGAN [40]	33.56
	Ours	**66.44**
Artistic style transfer	StyTr^2 [17]	20.89
	CCPL [75]	21.44
	Ours	**57.67**

Table S1. User study with 90 participants.

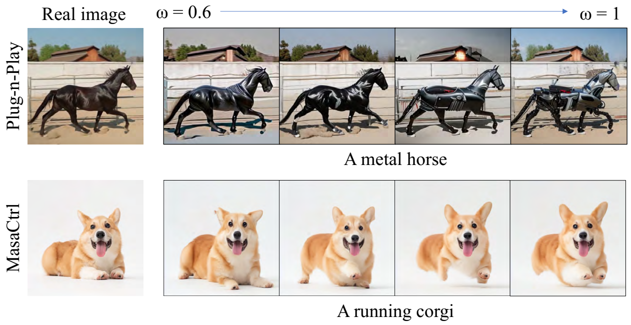


Figure S5. **Utilizing latent calibration to other methods.**. Increasing ω reflects injected results stronger when using other methods. For Stable Diffusion, we only use $\omega > 0.6$.

D.2. Comparison with the other methods.

Table S1 presents the results of a user study conducted with 90 participants to compare our method with existing methods. The participants were asked a question: "Which image is more natural while faithfully reflecting the original image and the content image?". We randomly selected ten images for content injections and thirty images for style transfer without any curation. The example images are shown in Figure S14-S16 (located at the end for clear spacing). Even though InjectFusion works on pretrained diffusion models without further training for the task, our method outperforms the others. We selects the recent methods from the respective tasks for comparison.

Although content injection does not define domains of images, it resembles image-to-image translation in that both

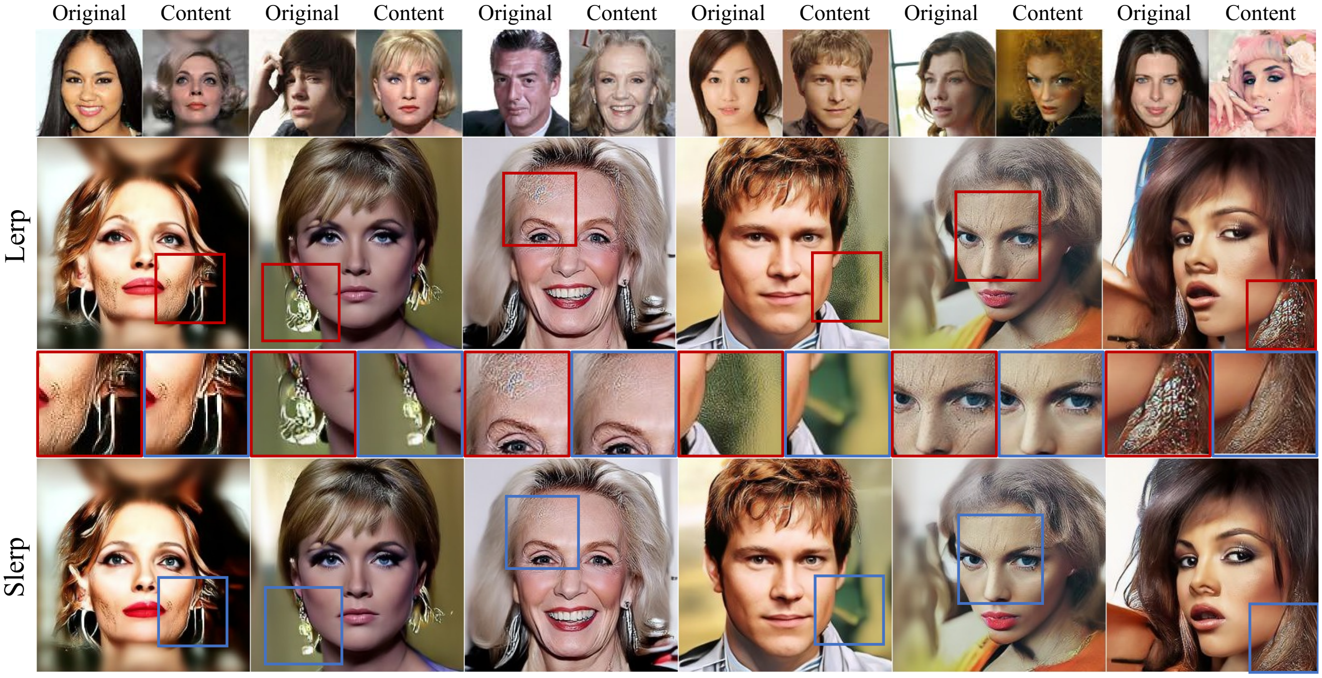


Figure S6. **Comparison between Slerp and Lerp.** Slerp reduces artifacts and distortions in Lerp. Note that We do not use other techniques such as quality boosting to evaluate the effect of Slerp only.

of their results preserve content of input images while adding different elements. Therefore, we show the differences between InjectFusion and those works in Figure S11. The resulting image of InjectFusion well reflects overall color distribution, color-related attributes (e.g. makeup), and non-facial elements (e.g. long hair, bang hair, decorations on a head) of the original images. Ours also reflect facial expression, jawline, and overall pose of the content image. On the other hand, the other works do not accurately reflect color-related attributes from the original images and also ignore fine-grained detail or spatial structure of the original image. They focus on preserving the structure of the content image.

D.3. Comparison with DiffuseIT

We provide more qualitative comparison with DiffuseIT [42] which uses DINO ViT [7]. As shown in Figure S17, InjectFusion shows comparable results without extra supervision. InjectFusion is highly proficient at accurately and authentically reflecting the color of the original image while avoiding artificial contrast, especially when there is a significant difference in color between the content and the original image (e.g., black and white). In contrast, DiffuseIT may not be able to fully capture the color of the original image in these scenarios. This discrepancy is due to the starting point of the reverse process. DiffuseIT utilizes the inverted $\boldsymbol{x}_T$ of the content image to sample and manipulate noise to match the target original image. The large gap in color distribution between the content and original images makes it challenging for DiffuseIT to overcome this difference entirely. Conversely, InjectFusion initially samples from the inverted $\boldsymbol{x}_T$ of the original image, making it easier to maintain the color of the original image. The original image is preserved through the skip connection.

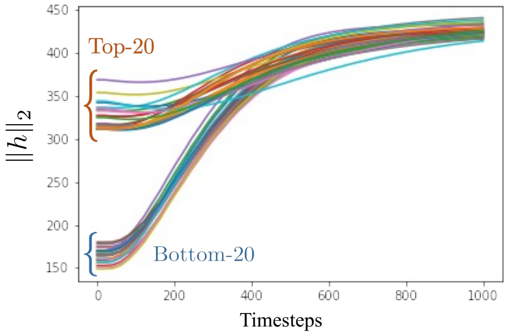


Figure S7. We choose $\boldsymbol{h}_t$ from the top 20 and bottom 20 samples in their norms among 500 samples. Each line represents a trajectory of $||h||_2$ during the reconstruction of a sample.

E. More analyses of Slerp

E.1. Comparison with Lerp

The intuition behind using Slerp is that we should preserve the correlation between $\boldsymbol{h}_t$ and its matching skip con-

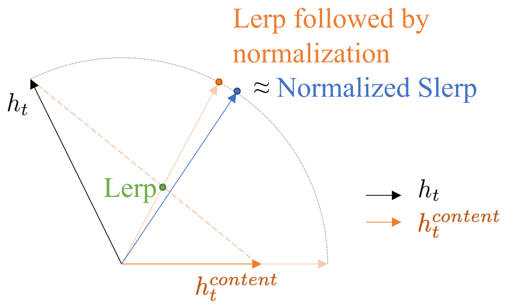


Figure S8. Visual comparison of Slerp and Lerp. The larger difference in norms of $\boldsymbol{h}_t$ and $\boldsymbol{h}_t^{content}$ leads to a larger gap between the results. Lerp followed by normalization is closer to Slerp than Lerp.

nection (§ 3.2). Here, we explore an alternative: Lerp. When $\boldsymbol{h}_t$ and $\boldsymbol{h}_t^{content}$ have different norms, using Lerp results in more artifacts in the final image as shown in Figure S6. This difference in norms of $\boldsymbol{h}_t$ is reported in Figure S7. Figure S8 illustrates the difference between Slerp, Lerp, and Lerp followed by normalization. Lerp may change the norm of $\mathbf{f}(\boldsymbol{h}_t, \boldsymbol{h}_t^{content}, \gamma)$ when the norm of $\boldsymbol{h}_t$ and $\boldsymbol{h}_t^{content}$ are different, leading to a decrease in image quality. However, Lerp followed by normalization produces results similar to Slerp. Still, we choose Slerp because it is easier to implement and less prone to errors.

E.2. Cumulative content injection

In addition to improving the quality of images, our approach allows us to control the amount of content injection by adjusting the $\boldsymbol{h}_t$-to-$\boldsymbol{h}_t^{content}$ ratio through Slerp parameter γ_t. A small γ_t results in a smaller amount of content injection. As mentioned in § 3.1, preserving the $\boldsymbol{h}_t$ component improves quality. However, there is a trade-off between the content injection rate and quality, and therefore, the value of $\boldsymbol{h}_t$ needs to be constrained. Further experiments to determine the proper range of γ are discussed in § 4.1.

Note that the effects of Slerp are cumulative along the reverse process as the content injection at t affects the following reverse process in $[t-1, t_{\text{edit}}]$. We provide an approximation of the total amount of injected content as follows. Assuming that the angle between $\boldsymbol{h}_t$ and $\boldsymbol{h}_t^{content}$ is close to 0 and the results of content injection at t are directly passed to the next $\boldsymbol{h}$-*space* at $t-1$ without any loss, then

$$\tilde{\boldsymbol{h}}_t = (1-\gamma)\boldsymbol{h}_t + \gamma \boldsymbol{h}_t^{content} \approx f(\boldsymbol{h}_t, \boldsymbol{h}_t^{content}, \gamma)$$

and

$$\boldsymbol{h}_{t-1} \approx \tilde{\boldsymbol{h}}_t.$$

Along the reverse process, $\tilde{\boldsymbol{h}}_t$ is recursively fed into the next stage. After n content injections, we get

$$\tilde{\boldsymbol{h}}_{t-n} \approx (1-\gamma)^n \boldsymbol{h}_t + \gamma \sum_{i=1}^{n} (1-\gamma)^{i-1} h_{t-i}^{content}. \tag{13}$$

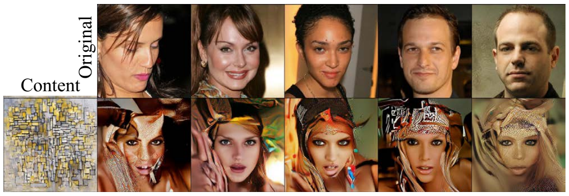


Figure S9. **Content image from unseen domain** Other than original images, $\boldsymbol{h}_t^{content}$ obtained from unseen domain results in poor images.

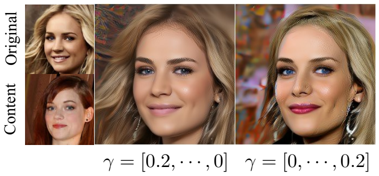


Figure S10. **Various interpolation ratio schedule.** γ is content injection rate.

As $0 \leq \gamma \leq 1$, the proportion of $\boldsymbol{h}_t$ decreases exponentially and the proportion of $\boldsymbol{h}_t^{content}$ accumulates during the content injection stage. It indicates that a large proportion of content is injected compared to γ of Slerp. For further details regarding the ablation study on γ, please refer to § 4.1.

F. Discussion details

As mentioned in § 5, Figure S9 shows that using out-of-domain images as content leads to completely distorted results. It implies that $\boldsymbol{h}_t$ cannot be considered a universal representation for all types of content.

Figure 12 shows the local mixing with various feature map mask sizes. Using the feature map mask, we can designate the specific area where the content injection is applied. Unfortunately, the $\boldsymbol{h}$-*space* has small spatial dimensions, limiting the resolution of the mask for local mixing.

G. γ scheduling

Figure S10 provides the results from alternative schedules. Gradually decreasing the injection along the generative process enhances realism, however, it may not accurately represent the content. Conversely, gradually increasing the injection better preserves the content but results in more artifacts. We keep the total amount of injection fixed in this experiment.

H. More related work

After [29,69] proposed a universal approach for Diffuson models (DMs), subsequent works have focused on controlling the generative process of DMs [1,8,14,21,39,41,44,49,50,58,72,76,77,80]. Especially, [4,43,57,71,81] have uncovered the role of intermediate feature maps of diffusion

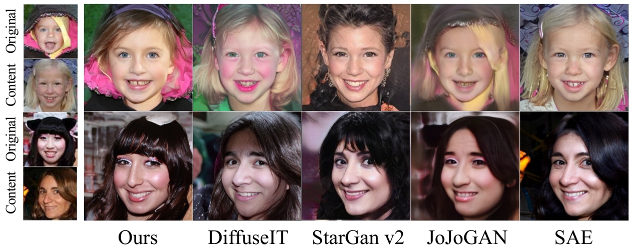
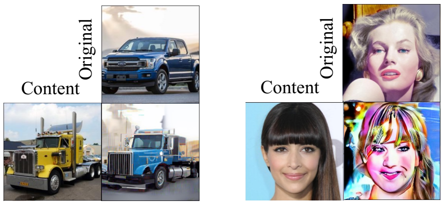


Figure S11. **More comparisons** InjectFusion shows different mixing strategy compared to the other methods.



(a) ImageNet (b) $Slerp(g_t, g_t^{content})$

Figure S12. (a) InjectFusion works on ImageNet. (b) Skip connection injection does not provide meaningful results.

models and utilized it for image editing, segmentation, and translation. However, we are the first to analyze the role of the latent variables $\boldsymbol{x}_t$ in DMs and apply it to content injection.

The research on controlling the generative process has been done in other generative models such as GANs [25]. [22, 31] introduce style transfer and image-to-image translation with GANs and there have been a number of works that focused on the style of images [2, 10, 11, 30, 55, 73, 78]. After StyleGAN [33, 36, 37], more diverse methodologies have been proposed [10, 12, 37, 40, 40]. However, most of them require training.

I. Stable diffusion experiment details

We provide more details of experiments with Stable diffusion. In Figure 13, we use conditional random sampling with Stable diffusion v2. In order to apply InjectFusion on Stable diffusion, there are 3 options with conditional guidance. 1) content injection only with unconditional output, 2) content injection only with conditional output, 3) content injection with both conditional/unconditional outputs. We find that using only the unconditional output for content injection resulted in poor outcomes, while the other two options produced similar results. Thus, we use only the conditional output for content injection in Figure 13.

Moving on to the implementation details for Stable diffusion, we set the scale to 9.0, use 50 steps for DDIM sampling, and employ the following prompts: for an original image, “a highly detailed epic cinematic concept art CG render digital painting artwork: dieselpunk steaming robot” and for a content image: “digital painting artwork: a cube-shaped robot with big wheels”, for an original image: “8k, wallpaper car” and for a content image: “concept, 8k, wallpaper sports car, ferrari bg”, for an original image: “a realistic photo of a woman.” and for a content image, “a realistic photo of a muscle man.”, original image: “A digital illustration of a small town, 4k, detailed, animation, fantasy” and for an original image: “A digital illustration of a dense forest, trending in artstation, 4k, fantasy.”

J. Definition of content

We provide more details of content definition used in § 4.1. We classify each of the attributes to determine whether they are from the content image or the original image by CLIP score (CS);

$$\text{CLIPScore}(x, a) = 100 * \text{sim}(\mathbf{E_I}(x), \mathbf{E_T}(a)), \tag{14}$$

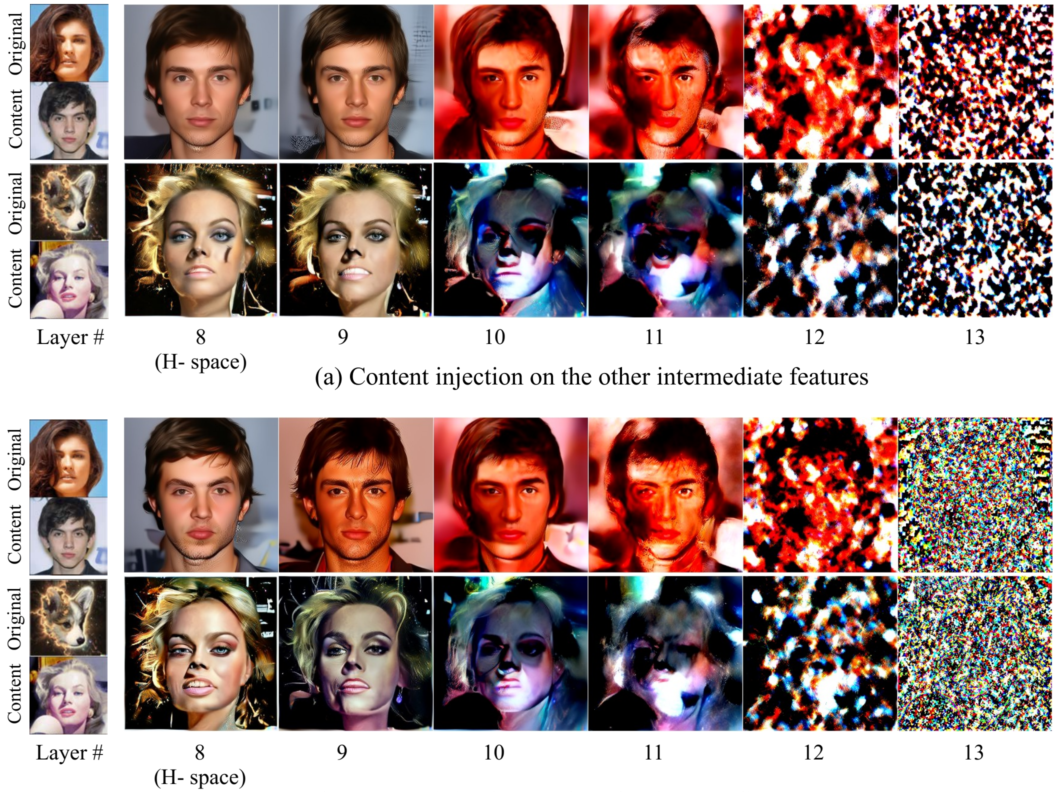
where x is a single image, a is a given text of attribute, $\text{sim}(*, *)$ is cosine similarity, and $\mathbf{E_I}$ and $\mathbf{E_T}$ are CLIP image encoder and text encoder respectively.

First, we calculate the CS between the desired texts and images, original image x_o, content image x_c, and result image x_r. Then, if the $|\text{CS}(x_o, a) - \text{CS}(x_r, a)| > |\text{CS}(x_c, a) - \text{CS}(x_r, a)|$ then we regard the attribute is from the content image and vice versa.

In order to ignore the case that x_o and x_c have similar attributes, the classified result was ignored when the difference between the two values was very small. Formally, if $\| |\text{CS}(x_o, a) - \text{CS}(x_r, a)| - |\text{CS}(x_c, a) - \text{CS}(x_r, a)| \| < \lambda_{th}$, we pass that sample for that attribute. We use 5k images and set $\lambda_{th} = 0.2$.

The result shows that content includes glasses, square jaw, young, bald, big nose, and facial expressions and the remaining elements include hairstyle, hair color, bang hair, accessories, beard, and makeup.

For the user study, we show the resulting image and ask people to choose the content or original image for each attribute. We use randomly chosen 100 images and aggregate the responses from 50 participants.



(a) Content injection on the other intermediate features

(b) Content injection on the other intermediate features with skip connection interpolation

Figure S13. The importance of h-space. When we inject features into additional layers, the results are disrupted. It supports h-space has semantic information and is the reason why we inject features into only h-space.

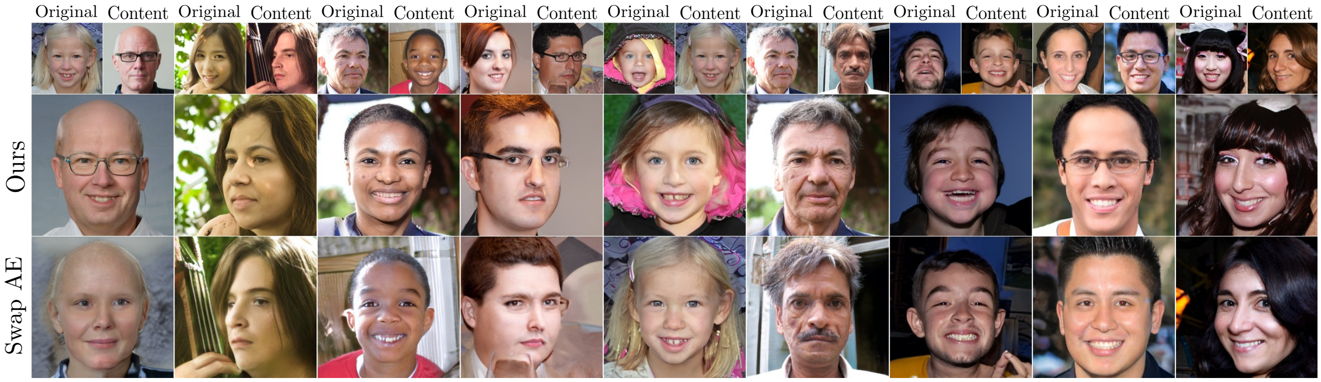


Figure S14. **Qualitative comparison of content injection on FFHQ.** InjectFusion is shown to be effective in reflecting content elements while preserving the overall color distribution of the original image.

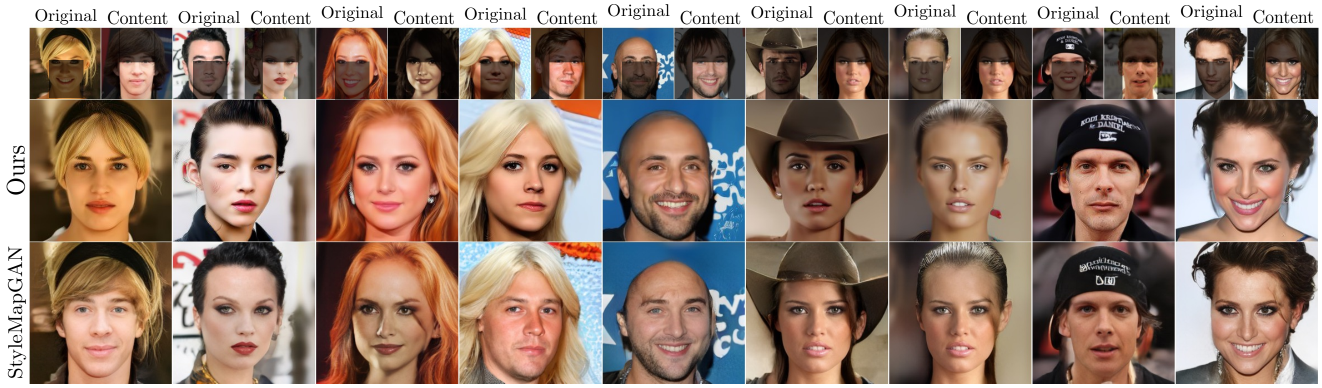


Figure S15. **Qualitative comparison of local mixing on CelebA-HQ.** Despite providing StyleMapGan with detailed segmentation guidance, there are noticeable artifacts in the resulting images, especially at the border lines of the mask. Furthermore, due to the differences in pose between the content and the original images, StyleMapGan struggles to seamlessly integrate the two images, resulting in less-than-optimal outcomes.

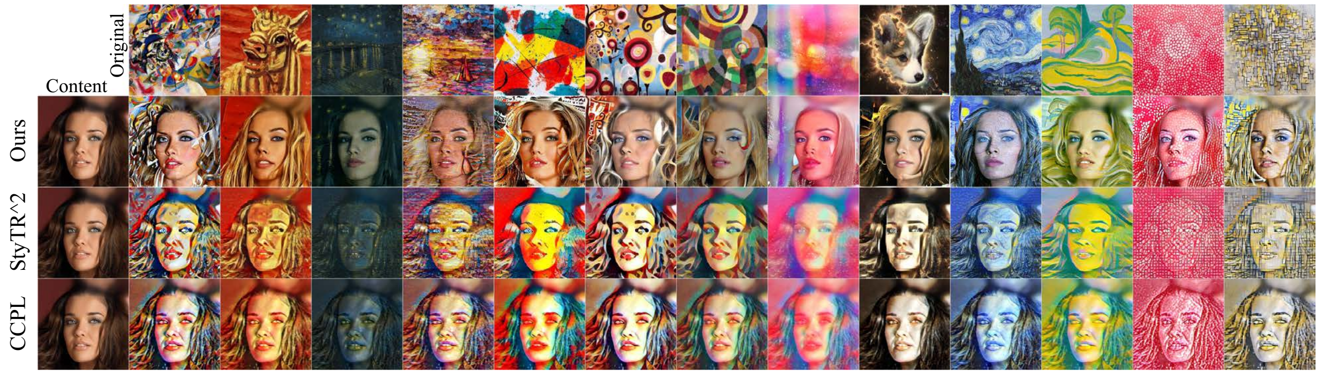
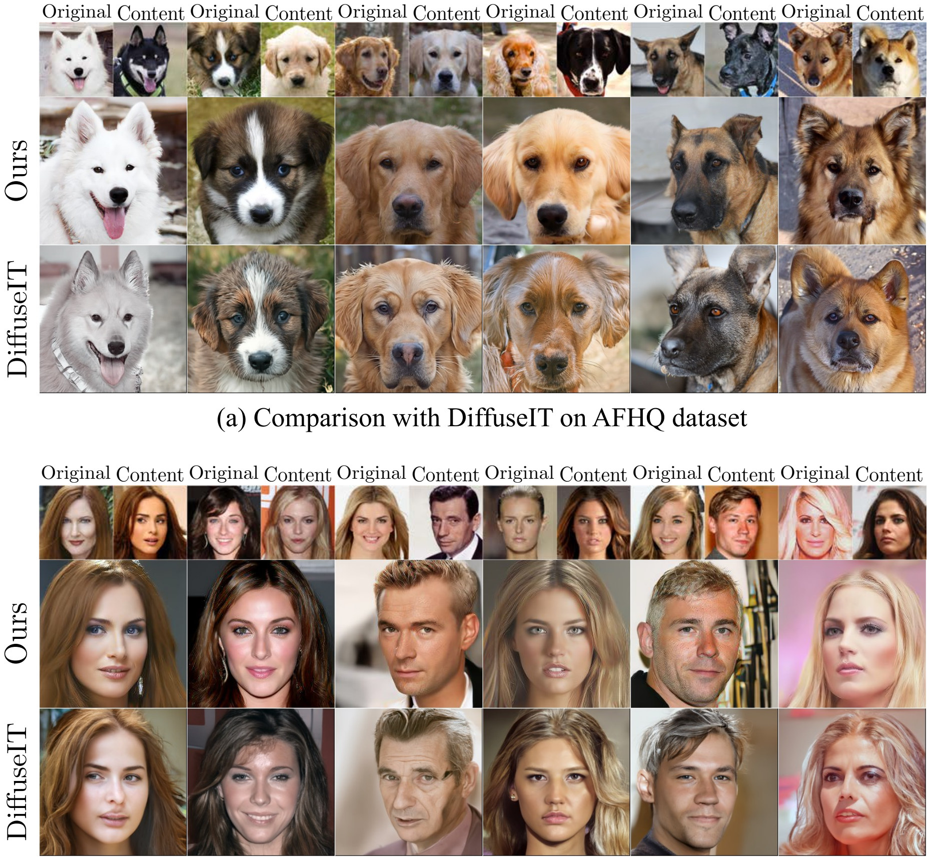


Figure S16. **Qualitative comparison between InjectFusionand style transfer methods with artistic references on CelebA-HQ.** InjectFusion allows using images from unseen domains as the original images, enabling the target content can be reflected on the artistic references. InjectFusion produces a harmonization-like effect without severe content distortion. Some high-level semantic color patterns of the original images are better reflected by InjectFusion than the others.



(a) Comparison with DiffuseIT on AFHQ dataset

(b) Comparison with DiffuseIT on CelebA-HQ dataset

Figure S17. **More qualitative comparison with DiffuseIT.** InjectFusion excels in fully and naturally reflecting the original color without creating artificial contrast, particularly when there is a significant gap between the content color and the style color (e.g., black and white). In contrast, DiffuseIT may not fully capture the original color in such cases.

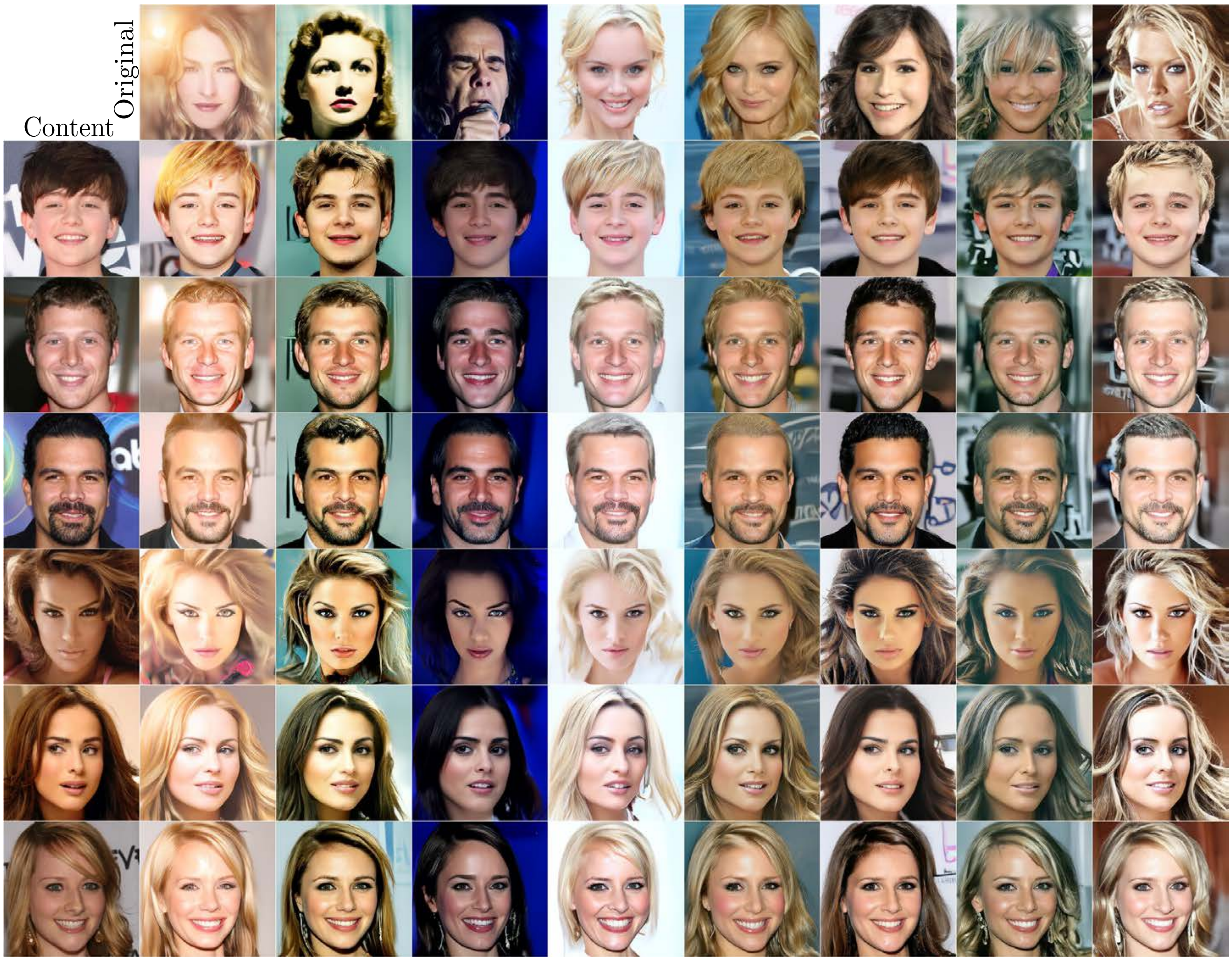


Figure S18. Qualitative results of content injection on CelebA-HQ.

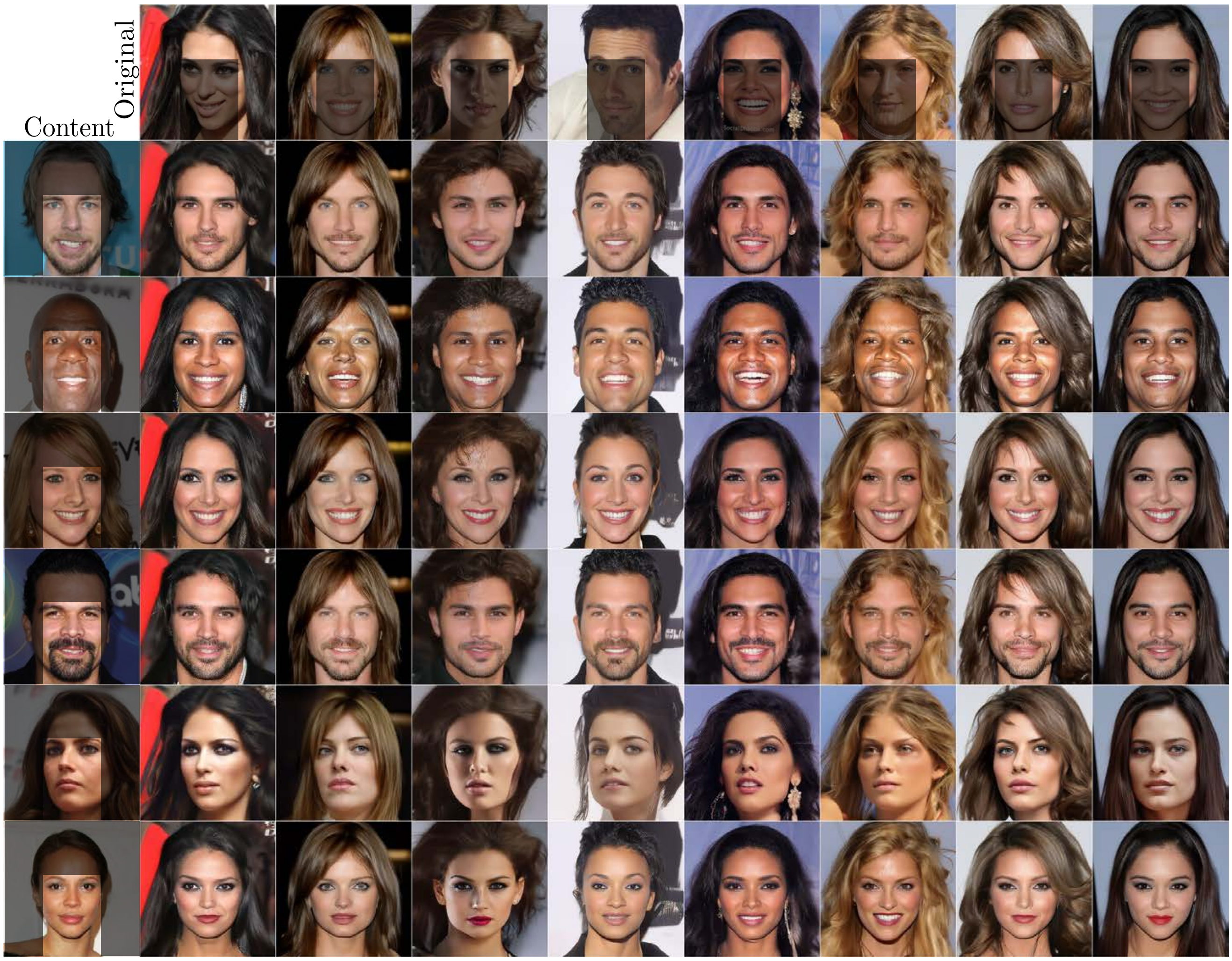


Figure S19. Qualitative results of local editing on CelebA-HQ.

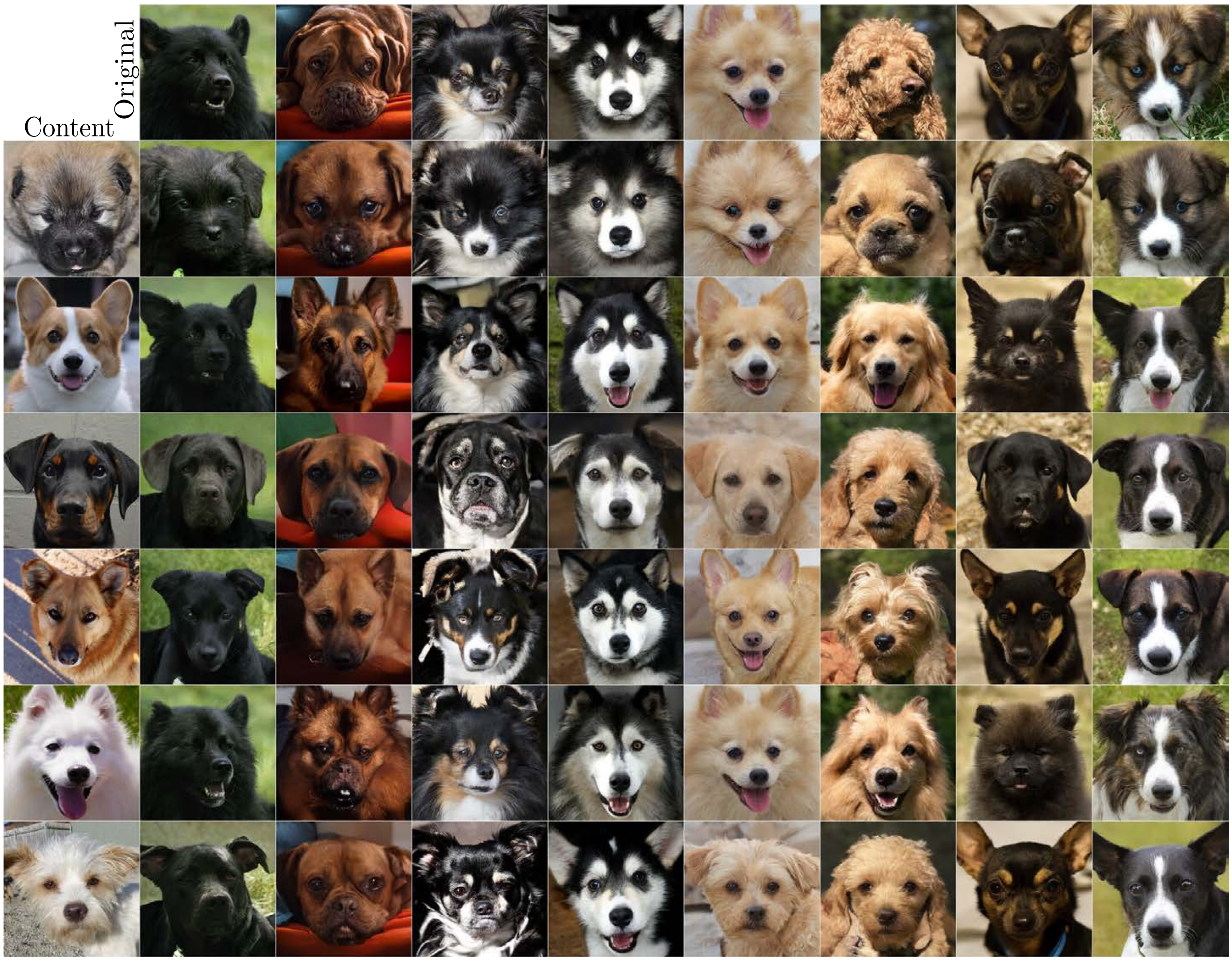


Figure S20. Qualitative results of content injection on AFHQ.

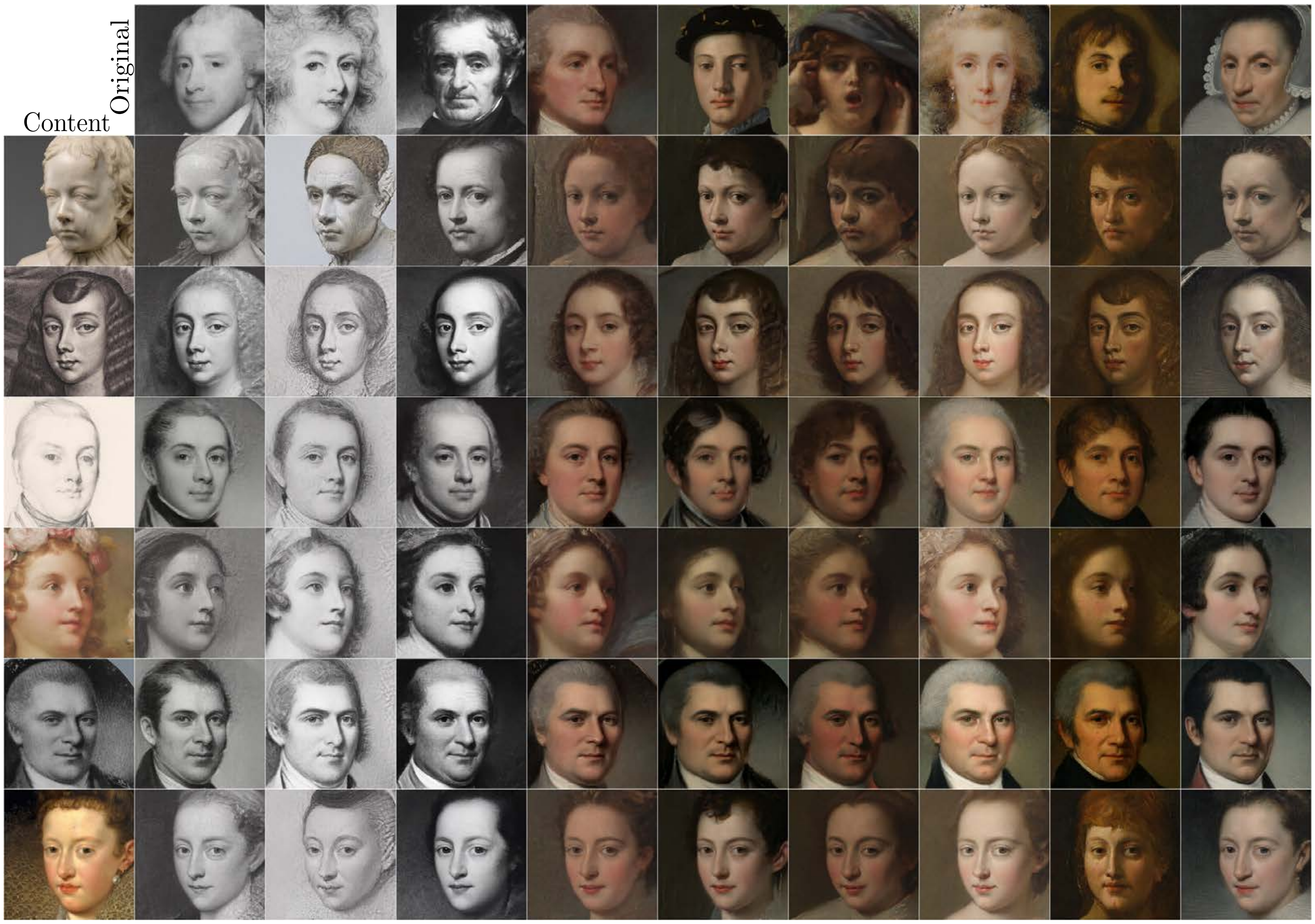


Figure S21. Qualitative results of content injection on METFACES.

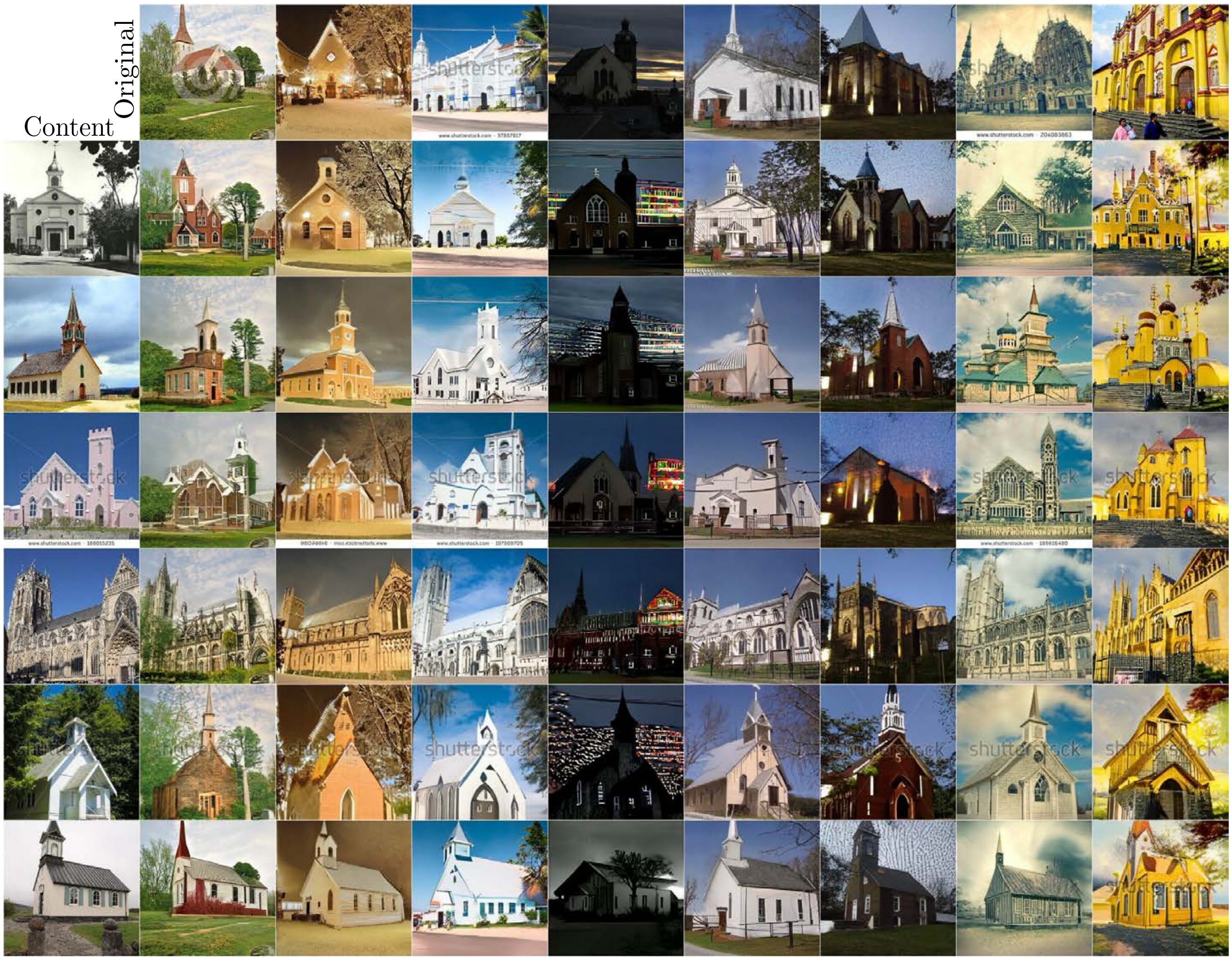

Original
Content
shutterstock
www.shutterstock.com · 57887917
www.shutterstock.com · 204083863
www.shutterstock.com · 166015235
www.shutterstock.com · 107303705
www.shutterstock.com · 185936480

Figure S22. Qualitative results of content injection on LSUN-church.

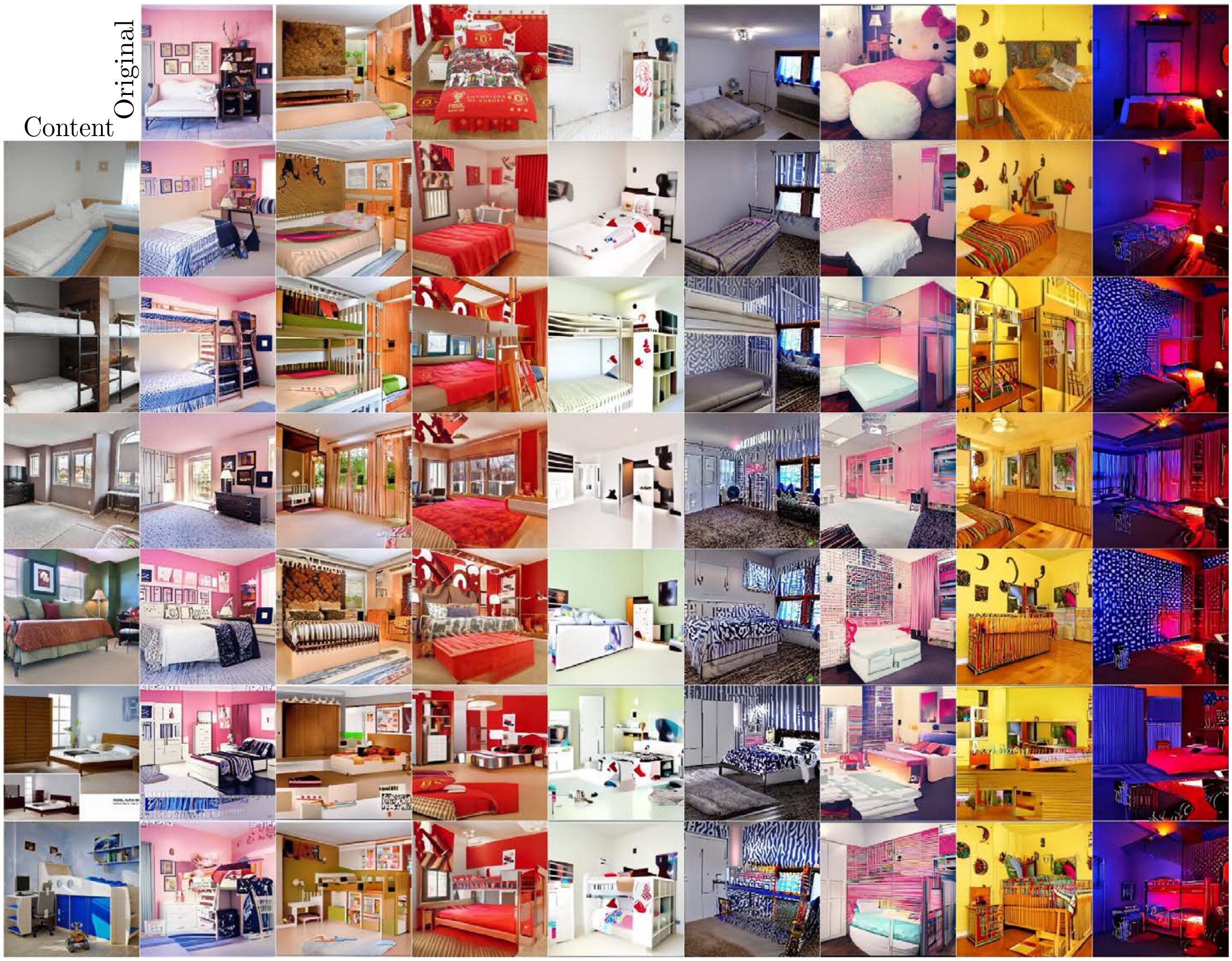


Figure S23. Qualitative results of content injection on LSUN-bedroom.

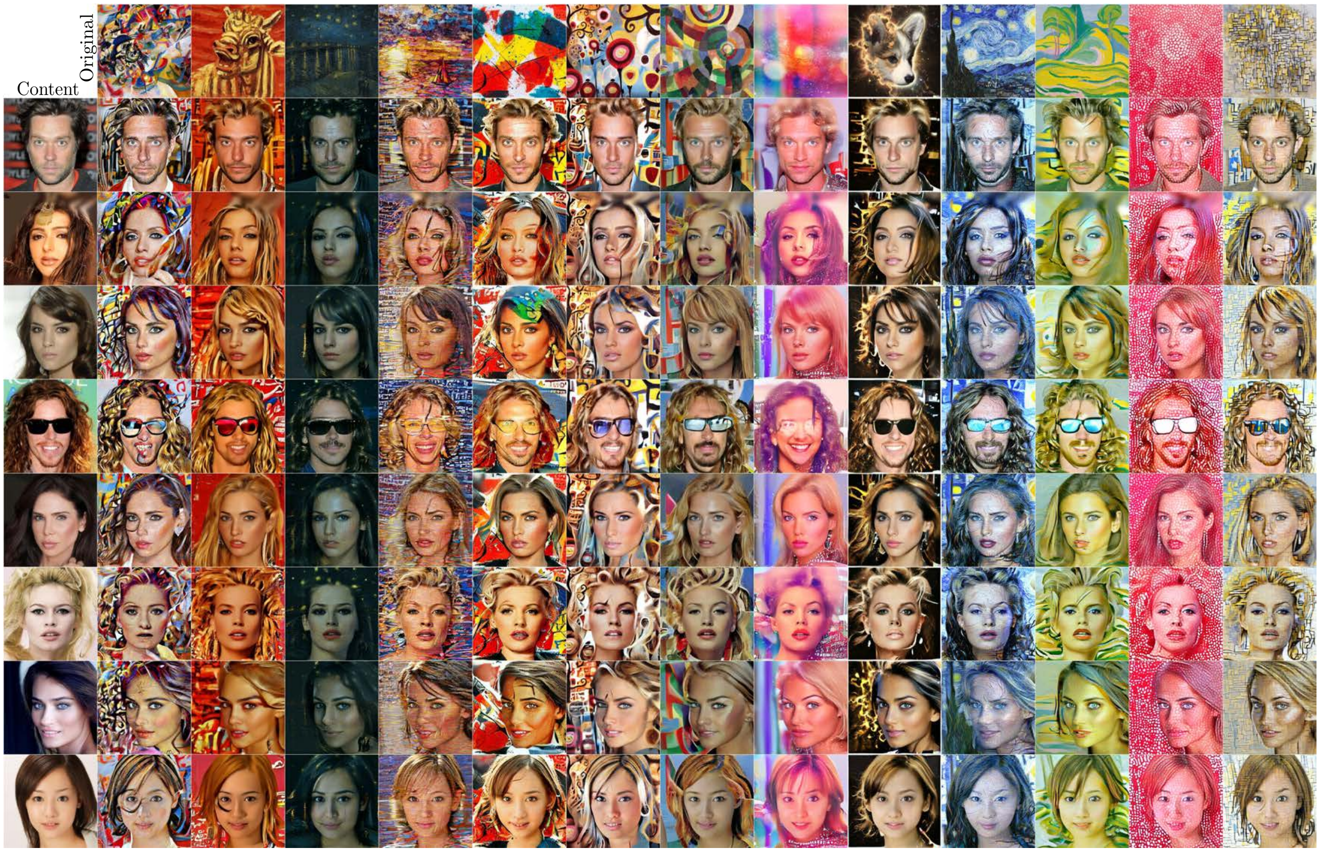


Figure S24. Qualitative results of content injection into artistic references with CelebA-HQ .