## A. Analysis of Condition Independence

Although integration of multiple guidances in the denoising process is popular recently in many tasks[], independence of conditions is ignored now. We first provide the total proof of Eq. (6).

**Property A.1.** *Denoting a sets of independent conditions as* $C = \{c_1, \cdots, c_k\}$ *with manually defined intensities* $w = \{w_1, \cdots, w_k\}$*, we can use multiple UNets* $\epsilon_\theta$ *to predict conditional and unconditional:* $\epsilon_{\theta_0}(z_t) = \nabla_z \log p(z_t)$ *and* $\epsilon_{\theta_i}(z_t, c_i) = \nabla_z \log p(z_t|c_i)$*. The reverse process in each timestamp is as follows:*

$$\hat{\epsilon}(z_t, C) = \epsilon_{\theta_0}(z_t) + \sum_{i=1}^{K} w_i(\epsilon_{\theta_i}(z_t, c_i) - \epsilon_{\theta_0}(z_t)) \quad (6)$$

*Proof.* Since $C = \{c_1, \cdots, c_k\}$ are independent, we have $p(C|z_{t+1}) = \prod p(c_i|z_{t+1})^{w_i}$. The denoising process is then converted into:

$$\begin{aligned}
&\nabla_{z_t} \log p(z_t) + \nabla_{z_t} \log p(C|z_t) \\
=&\nabla_{z_t} \log p(z_t) + \sum_{i=1}^{K} w_i \nabla_{z_t} \log p(c_i|z_t) \\
=&\nabla_{z_t} \log p(z_t) + \sum_{i=1}^{K} w_i \nabla_{z_t} \log \frac{p(z_t|c_i)}{p(z_t)} \\
=&\nabla_{z_t} \log p(z_t) + \sum_{i=1}^{K} w_i (\nabla_{z_t} \log p(z_t|c_i) - \nabla_{z_t} \log p(z_t))
\end{aligned}$$

In Stable Diffusion, we use a UNet $\epsilon$ to predict each term: $\epsilon(z_t) = \nabla_z \log p(z_t)$ and $\epsilon(z_t, c_i) = \nabla_z \log p(z_t|c_i)$. □

An important assumption is the independence of each condition. Here we provide a simple example to reveal the condition independence and illustrate how we can utilize multiple conditions in in-domain generation. We use two text conditions $c_1$ =`a photo of a cat` and $c_2$ =`a photo of a dog`, and attain the guidance by: $\hat{\epsilon} = (1 - 2w)\epsilon(z_t) + w \cdot \epsilon(z_t, c_1) + w \cdot \epsilon(z_t, c_2)$. The expected result, which was to produce one cat and one dog, did not occur; instead, a hybrid creature, combining features of both a cat and a dog, was generated as can be seen in Fig. 13.

The theoretical explanation for this phenomenon is that during the denoising process, the denoising direction is independently guided by each component to maximize the likelihood of each condition independently. It means that the generative result achieves the high probability of both $p(c_1|x)$ and $p(c_2|x)$. Hence, an effective way to utilize multiple guidance for synthesis is by using conditions which can represent the whole image. For example, in text-guided face-domain generation, using `a photo of face, wearing glasses` would be much better than using `wearing glasses`.

Figure 13. **Example to reveal condition independence.** We use two text conditions `a photo of a cat` and `a photo of a dog`. The generative results represent a mixture creature of dog and cat.

## B. Condition Decoupling


Figure 14. **Example of condition decoupling.** Decouple conditions can make contents independent. It helps mitigate the bias in diffusion priors, leading to more diverse results.

Apart from domain guidance and control guidance, we further demonstrate a general condition decoupling technique. It can also explain that we could using our in-domain diffusion model to generate out-of-domain results (like stylization generation). Using multiple guidances also has the unique ability to decouple relationships between contents that are typically related in the real world, thereby enabling the generation of more diverse images. For instance, if the goal is to generate images of a sunflower in the style of Van Gogh, as opposed to replicating Van Gogh's specific sunflower paintings, we can apply two conditions: `sunflower` and `in Van Gogh style`. This distinction is showcased in Fig. 14. Using the combined prompt


Figure 15. **Ablation study on 3D generation.** During the 3D fine-tuning process, we replaced our porcelain model with basic SD, resulting in a noticeable decline in generation quality. The distinctive porcelain patterns in the output disappeared completely.

`sunflower in Van Gogh Style` tends to produce images closely resembling Van Gogh's paintings. In contrast, employing decoupled conditions results in a much broader diversity of images. This method proves equally effective for generating concepts similar to those in the training dataset, such as `rose`.

# C. Experimental Details

## C.1. Inference and Evaluation

We employ the DPM-Solver++ scheduler [28] with 100 steps across all models while using 25 steps for stylized results to achieve better results. Regarding the guidance scale, we perform grid searches on all models and conditions, with an interval of 0.25. In text-guided generation, we combine the prompts for baselines, like `a photo of <V> face, wearing glasses`. To quantitatively measure the fidelity and controllability of the generated images, we incorporate the FID in unconditional generation, and alignment in conditional generation. Moreover, we also undertake a user study to gather human preferences, providing a qualitative dimension to our evaluation criteria and ensuring a holistic view of the performance of the methods under scrutiny

**Text-guided Generation.** Although the measure of text-image similarity, often quantified by CLIP scores, is a common metric for evaluating the alignment in text-guided image generation, we observed its limitations in accurately reflecting facial attributes. For instance, in the case of the prompt 'wearing glasses,' the difference in CLIP scores between images with and without glasses is relatively marginal, often around a value of 1 (e.g., 18 to 19). Due to this lack of pronounced differentiation, we adopt an alternative approach for evaluation.

We utilize a set of attribute-specific prompts, detailed in Tab 2, to generate images. Subsequently, we employ facial attribute predictors to ascertain whether the specified attribute (e.g., glasses) is accurately represented in the generated images. The effectiveness of our method is then quantified by the ratio of successful samples where the controlled attribute is correctly manifested in the results.

Table 2. **Prompts used for text-guided generation evaluation.** We generate 100 images of each prompt for evaluation.

| | |
|---|---|
| `wearing glasses` | `wearing sunglasses` |
| `wearing hat` | `smilling` |
| `male` | `female` |
| `white people` | `black people` |
| `asian people` | `square face` |

## C.2. Fine-Tuning Baselines

To establish a comprehensive baseline, we employ several state-of-the-art fine-tuning methods: Textual Inversion [13], DreamBooth [41], Custom Diffusion [22], and OFT [34], along with our method. Textual Inversion and Custom Diffusion serve as parameter-efficient comparisons, streamlining the experimental setup, while OFT is included for its regularization properties. DreamBooth fine-tunes a concept token `<V>` and UNet's parameters without class-specific prior preservation loss since it is not compatible to customize concept. The experimental distinction between baselines and our concept-centric diffusion models lies in the use of text prompts. We adopt prompt templates inspired by DreamBooth [41], utilizing formats such as `a photo of <V> face`. These additional tokens `<V>` are omitted in compared methods where text embedding is not trained

**Spatial-guided Generation.** To assess the effectiveness of spatial-guided generation, we select a sample of 200 images from the CelebA-HQ dataset. For each image, we generate corresponding canny images to use as conditions. We then produce 5 generated results per condition and compute the discrepancy between the canny conditions and the corresponding generative results, providing a quantitative measure of the alignment accuracy.

**User Study.** To gauge human preferences in both unconditional and conditional generation contexts, we organize a user study. We gain generative results from all methods using the same random seed across all methods and present these images to participants in a random sequence. Participants are instructed to choose the best image from the given set. We record the frequency with which each method is selected as the best, and this data is used to calculate the win rate for each method. The win rates serve as an indicator of human preference (denoted as 'Pref.') and are presented in the main paper.

### C.3. Other In-domain Generation Settings

#### C.3.1 Image Editing

We following SDEdit to edit facial images with diffusers implementation[1]. We use noising strength of 0.6 and inference steps of 20. We follow Sec. 3.4 to adjust the guidance scales. The editing process is the same as text-guided in-domain generation, we utilize original SD1.5 to predict unconditional guidance and text-conditioned guidance and use the trained facial domain diffusion model to predict domain guidance.

#### C.3.2 Text-to-3D Generation

Our design involves three stages to generate a 3D porcelain. First, we use our model to create a 2D porcelain image. Next, we employ the classic Zero-1-to-3 model for image-to-3D conversion to obtain a rough target 3D model. Finally, we fine-tune this model using our porcelain model and SDS loss. During the fine-tuning process, we set the CFG parameters to 100 on SD and 50 on our porcelain model. Additional results is shown on Fig 18. Our implementation is based on stable-dreamfusion[2].

**Effects of our porcelain model in 3D generation.** During the 3D fine-tuning process, we replaced our porcelain model with basic SD, resulting in a noticeable decline in generation quality. The distinctive porcelain patterns in the output disappeared completely, as shown in Fig 15.

## D. More Qualitative Results

### D.1. Unconditional Concept-centric Generation

- Fig 16 illustrates the comparison of unconditional generative results on FFHQ.
- Fig 17 illustrates the comparison of unconditional generative results on AFHQv2.

### D.2. Conditional Generation

Fig 19 illustrates the comparison of text-guided generative results on FFHQ.

Fig 18 illustrates the results of 3D generation within porcelain domain.

## E. Limitations

Using multiple models to estimate different guidances slightly increases both inference time and memory usage. Originally, the generation process required predicting two guidances, but in-domain generation requires predicting three guidances, resulting in a 50% increase in computation time. To accelerate this process, we utilized multiple GPUs by placing the concept diffusion model and the original diffusion model on separate GPUs for parallel computation, which only increased the generation time by approximately 16%. Regarding memory usage, incorporating an additional concept diffusion model increases the memory requirement for 512-resolution generation from 3.5GB to 5GB, which is acceptable for most GPU devices. One foreseeable solution to the computation time and memory issues is to train on a distilled, smaller version of Stable Diffusion, as learning domain guidance does not require a large UNet. We plan to explore this in future work. Meanwhile, text prompts can not be too much out-of-domain, which causes conflict between domain guidance and control guidance.

[1] https://github.com/huggingface/diffusers/blob/main/src/diffusers/pipelines/stable_diffusion/pipeline_stable_diffusion_img2img.py

[2] https://github.com/ashawkey/stable-dreamfusion


Figure 16. **Unconditional generation results on FFHQ.** We illustrate the unconditional results of all models trained from our method and baselines.


Figure 17. **Unconditional generation results on AFHQv2.** We illustrate the unconditional results of all models trained from our method and baselines.

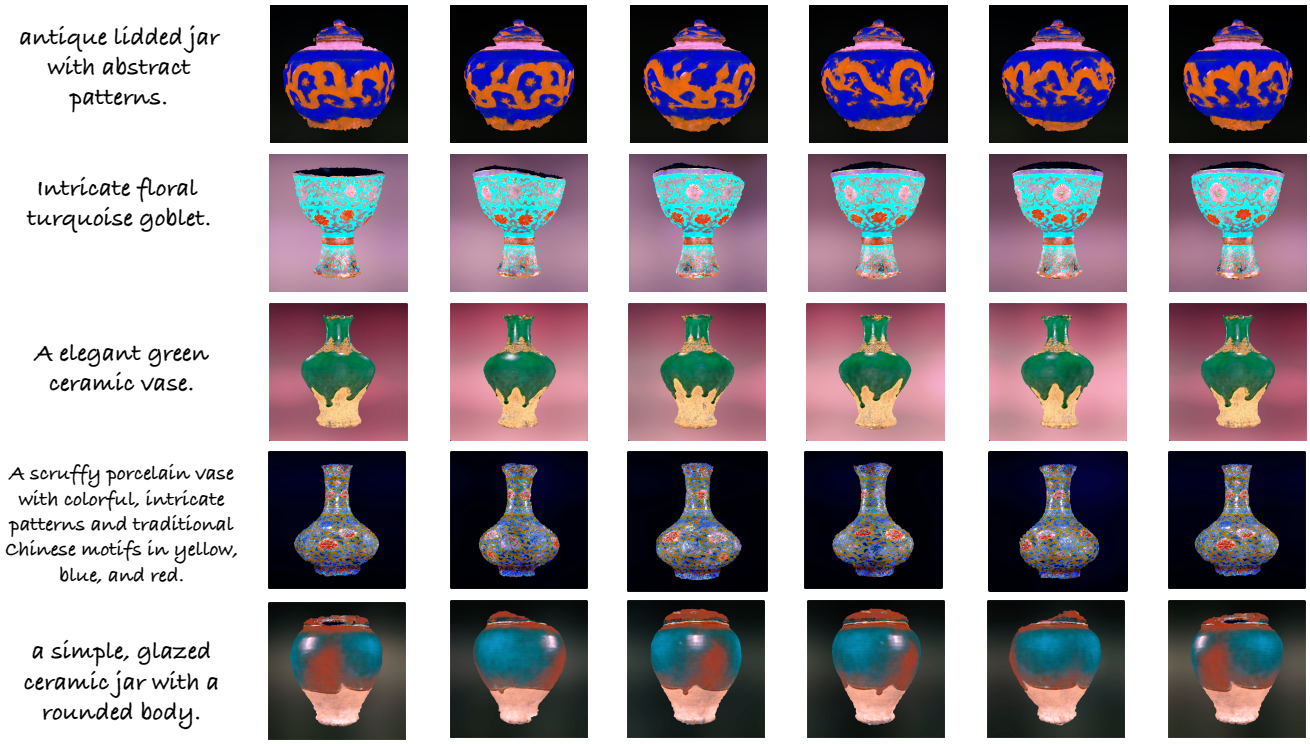


Figure 18. **Additional results of 3D generation with porcelain model.**

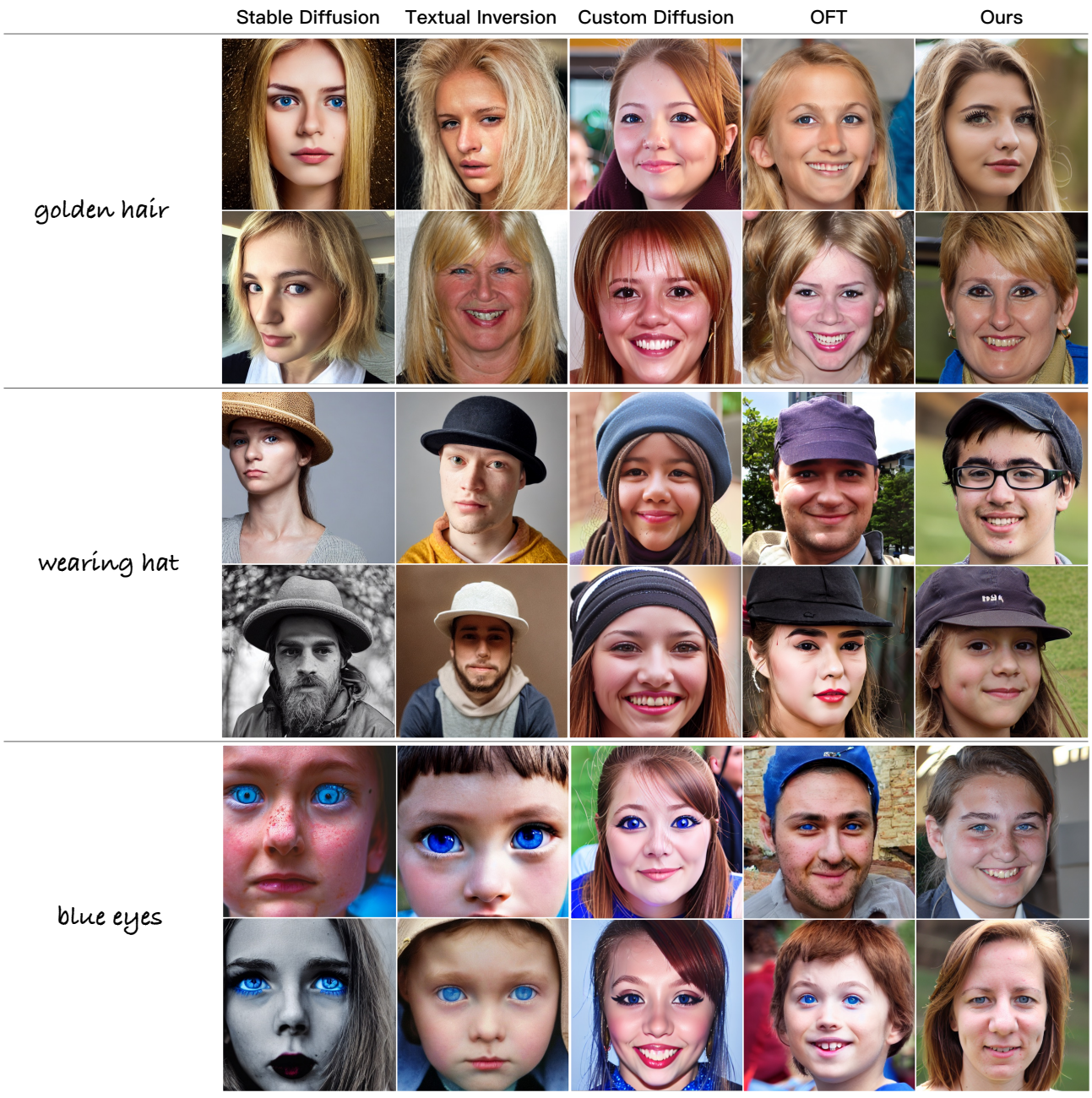


Figure 19. **Text-guided generation comparison across baselines and our method.** Since fine-tuning with large-scale training process almost loses controllability (as shown in Fig 11), we evaluate other methods in this part.