We sincerely thank all reviewers and appreciate the positive comments on "a solid design", "solving a key problem" and "comprehensive experiments". In the following, we address the concerns from each reviewer.

To Reviewer #1:

Q 1.1 Novelty. There are two key differences between DBT [10] and our work. 1) Different tasks. Our SariGAN aims to disentangle fine-grained semantics for unsupervised generative models. We achieve this by mining (with SGM) and advancing (with AdaGN) the intrinsic attributes of channels (i.e., semantics) in the latent space of relative importance (line 50-51 in the paper). While DBT [10] proposed to learn group bilinear features for classification. 2) Different grouping algorithms. DBT [10] is not designed for generation tasks. For example, the uniformly divided channels (per group) prohibit learning from more channels to represent complex semantics. Also, the hand-crafted block diagonal constraint propagates inconsistent gradient against the generation task. We have observed an 11.9% relative drop in terms of FID on LSUN CATS if using DBT [10] as grouping algorithms. We will add this discussion to related work.

For InfoGAN loss, it is a general framework designed for latent factors disentangling. We use InfoGAN in the paper for ensuring that inter/intra-group semantics can be well-disentangled. In Table 2, experiments demonstrate the gain of 16% achieved by using infoGAN loss. We believe such improvement is non-trival, and the using of InfoGAN is worth mentioning. We will take your helpful suggestions and make the statement more clear in the final version.

Q 1.2 The relation of kernels' similarity and feature channels' similarity. Given a well-trained model, the semantic of each feature channel is decided by the corresponding kernel. In experiments, we randomly generate 10k samples to calculate the pairwise similarity of feature channels and compare it with the similarity of kernels. As shown in Figure 1 (a), kernels' similarity and feature channels' similarity are positively correlated. Note that if we directly use feature channels, the similarity would be imprecise in each training batch due to the limited batch size (e.g., 32). Specifically, according to the law of large numbers, the more samples calculated, the more precise similarity approximated. As shown in Figure 1 (b), the similarity of channels calculated in a batch causes severe inconsistency. We will add discussions on this problem according to your kind suggestions.

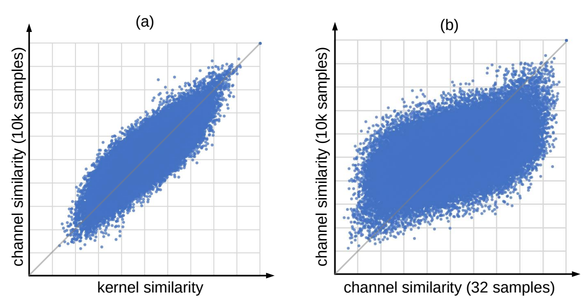

Figure 1: An illustration of the similarity consistency of kernels, feature channels, and feature channels in a batch. The more diagonally concentrated, the better consistency. It can be observed that the similarity of kernels in (a) is more consistent with feature channels than that of channels calculated in a batch in (b).

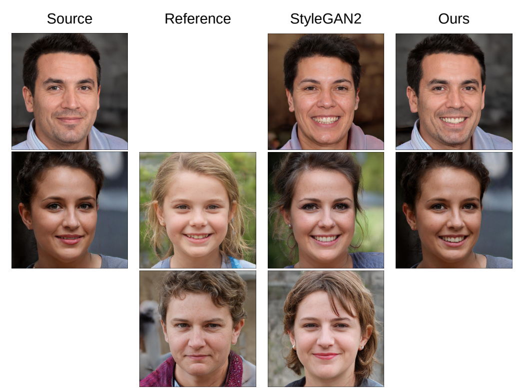

Figure 2: Qualitative comparison of StyleGAN2 and SariGAN. SariGAN can control a specific semantic (e.g., mouth) while preserving identities. StyleGAN2 makes several unexpected changes, as the semantics of mouth, eyes, and hairstyle are highly-entangled.

Q 1.3 Qualitative comparison. We provide qualitative comparisons with StyleGAN2 in Figure 2. It can be observed that SariGAN can achieve semantic-level controls (e.g., control the mouth), while StyleGAN2 can only achieve scale-level controls (e.g., the semantics of mouth, eyes, and hairstyle are still entangled). More cases for semantic-specific controls can be found in Figure 2 in our supplementary. Thanks for your valuable comments, and we will add more qualitative comparisons in the final version.

For inpainting results, we provide more cases in Figure 4 in the supplementary material to eliminate case biases, which show consistent qualitative improvements by SariGAN comparing over SOTA.

To Reviewer #3: Definition: Unconditional image generation (synthesis) is the task of generating new images unconditionally from an existing dataset. And image inpainting aims at filling missing pixels in a damaged image given a corresponding mask (line 220). The $_1$ and $_2$ in Equation 7 are set to be 2 and 10, respectively (line 191).

The inter-group and intra-group embeddings disentangle semantics in different levels. The former controls semantics like pose, age, and gender (Figure 4 in the paper); and the latter controls semantics like mouth, eyes, and glasses (Figure 2 in our supplementary). Thanks for your valuable suggestions, and we will make these clearer in the final version.

To Reviewer #4: Thanks for your valuable comments. The discriminator consists of 16, 18, and 20 layers for the CATS, CARS, and FFHQ datasets, respectively (i.e., two layers for each resolution 4^2 256^2=512^2=1024^2 and two additional layers). We will add all these details in the final version and release both codes and models.