

## 6. Ablations

To further motivate our design choices, we show additional ablation results.

**Reconstruction loss.** Specifically, we further compare using our method against using an identical setup as ours, but directly learning the hypernetwork through the task-specific loss. More formally this can be written as updating parameters  $\phi$  in the following equation:

$$\phi \leftarrow \phi - \eta \nabla_{\theta_{H_\phi(c,T)}} L_{task}(H_\phi(c,T), c). \quad (5)$$

While this looks like a good idea at first glance training results in over-adherence to the single example that was given—it results in outputting the same image regardless of the prompt, as shown in Fig. 12. We report the quantitative performance of optimizing using the task-specific loss, *i.e.*, the reconstruction loss in Tab. 3 for the CelebA dataset and in Tab. 4 for the AFHQ dataset. Note how the CLIP-T scores are much worse than with our method, which effectively shows what happens in Fig. 12.

**Fast fine tuning (FFT).** We further report our performance without the fast fine-tuning applied to our model. As shown, there is a slight decrease in performance in terms of CLIP-I and DINO, but both are still much higher in the case of CelebA than the baselines, and for AFHQ, still comparable. Note, however, how CLIP-T scores are higher, compared to any of the baselines, and even our method with FFT. This actually indicates better adherence to prompts, as demonstrated earlier in Fig. 3 and in Fig. 4. We further argue that low Face Recognition score in Tab. 3 is not a critical factor, as once they are transferred to different contexts, such as the funkopop figure in Fig. 3, it is natural that they are not high.

## 7. Extra Qualitative Results

We present additional results generated by our hypernetwork, both in its direct output form and after fast fine-tuning. Specifically, we include images from the AFHQ dataset [7] sampled directly from the hypernetwork (Figure 10) and after fast fine-tuning (Figure 11). Similarly, we show results for the CelebA dataset [25] directly from the hypernetwork (Figure 8) and after fast fine-tuning (Figure 9). These results demonstrate that our hypernetwork produces reasonable outputs without additional tuning, but higher-quality images can be achieved with the fast fine-tuning process.

**User Study.** Automatic face reconstruction metrics have inherent limitations, as they are primarily designed for *aligned, photorealistic human faces*. Consequently, their reliability diminishes significantly when evaluating stylized or abstract prompts, such as ‘as funkopop figure’ or ‘as a graffiti mural.’ To address this limitation, we supplement

the automatic metrics with a human preference survey, providing a more nuanced assessment of how well each method preserves the identity of the subject across diverse prompts.

In this user study, participants were asked to indicate their preferred image generation method, given the conditioning image and a specific prompt. Responses were collected from a total of 903 evaluations. The summarized preference scores are presented in Tab. 5, clearly indicating that our method is significantly favored compared to baseline approaches.

While automatic metrics are included for completeness, these human preference results offer a more reliable indicator of method effectiveness, particularly highlighting the robustness of our method in maintaining subject identity across stylized or abstract image generation scenarios.

Ablation	Face Rec.	CLIP-I	DINO	CLIP-T
Ours	<b>0.325</b>	<b>0.605</b>	<b>0.639</b>	0.268
Recon. loss	0.068	0.211	0.138	0.211
Ours w/o FFT	0.157	0.582	0.532	<b>0.284</b>

Table 3. **CelebA Ablations** – Our method provides significantly better performance than directly optimizing through the task-specific loss (the reconstruction loss). Most evidently, directly learning to reconstruct results in low adherence to prompts, as shown by the low CLIP-T scores. In fact, as shown in Fig. 12, the method starts ignoring the prompt. Fast fine-tuning helps, but is not critical.

Ablation	CLIP-I	DINO	CLIP-T
Ours	<b>0.664</b>	<b>0.807</b>	0.277
Recon. loss	0.607	0.070	0.201
Ours w/o FFT	0.495	0.746	<b>0.285</b>

Table 4. **AFHQ Ablations** – Similar to the CelebA ablations, our full model is shown to perform best. As with the CelebA experiments, the Reconstruction loss ablation initially learned to simply output a copy of the condition image before breaking completely.

Method	Preference (%)
Ours	<b>43.5</b>
DreamBooth	35.1
Textual Inversion	21.4

Table 5. User study results show that our method is preferred for identity preservation over benchmarks.



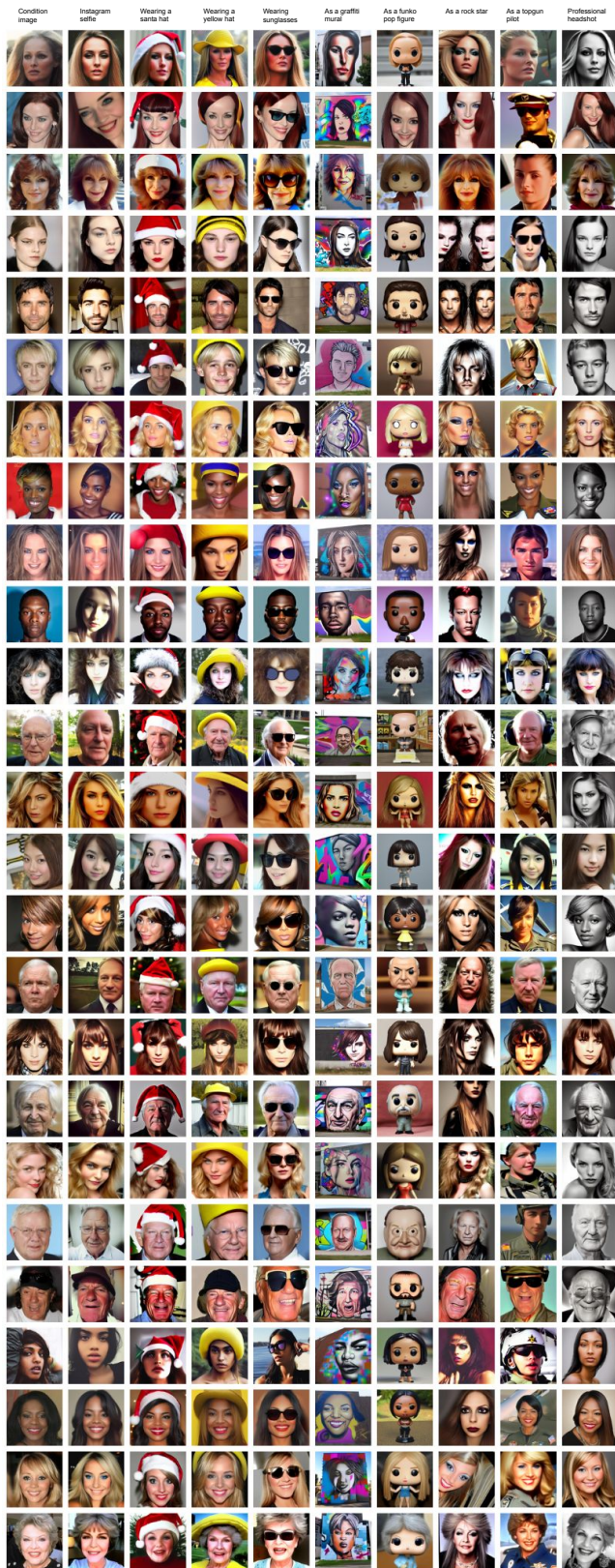


Figure 8. **Celeba without fast fine tuning** – Each of these images had their dreambooth parameters estimated in a single forward pass of our hypernetwork field. Our model can be seen to produce reasonable results even without extra fine tuning.

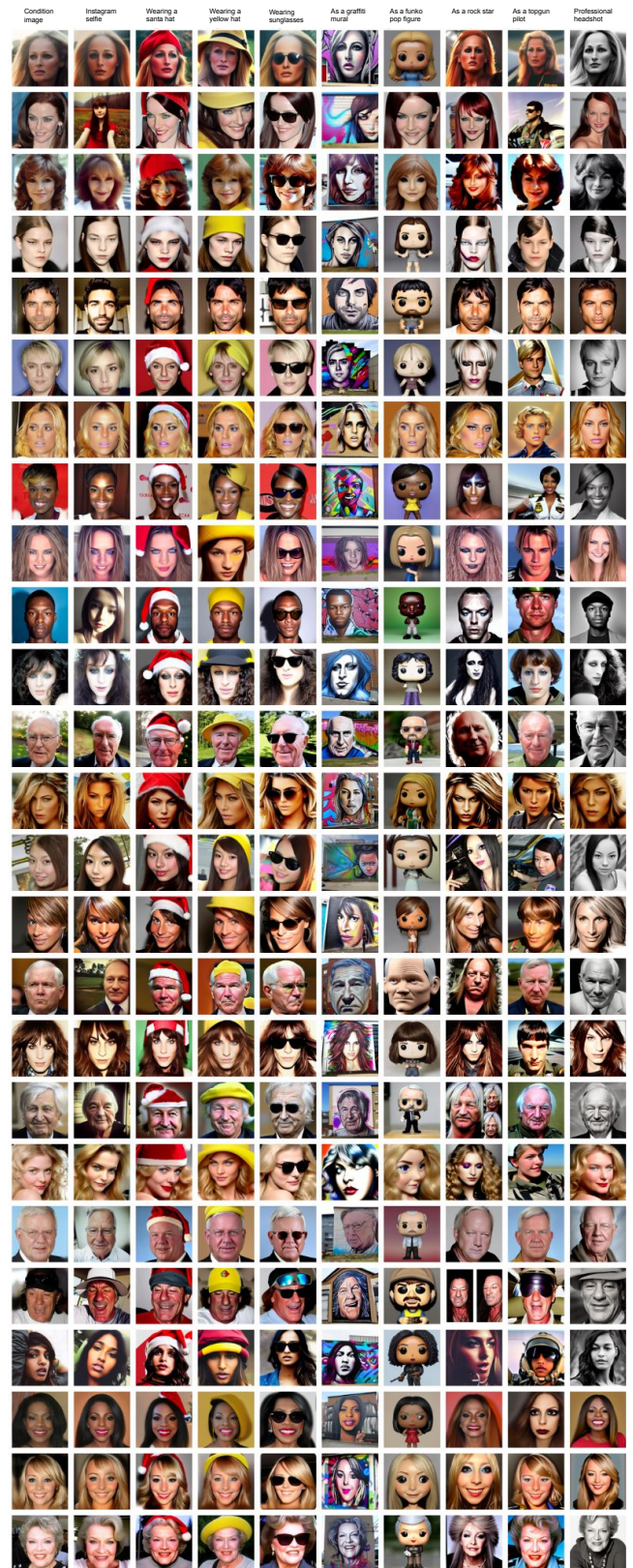


Figure 9. **Celeba with fast fine tuning** – Each of these images had 50 iterations of DreamBooth fast fine tuning applied to them.





Figure 10. **AFHQ without fast fine tuning** – Each of these images had their dreambooth parameters estimated in a single forward pass of our hypernetwork field. Our model can be seen to produce reasonable results even without extra fine tuning.

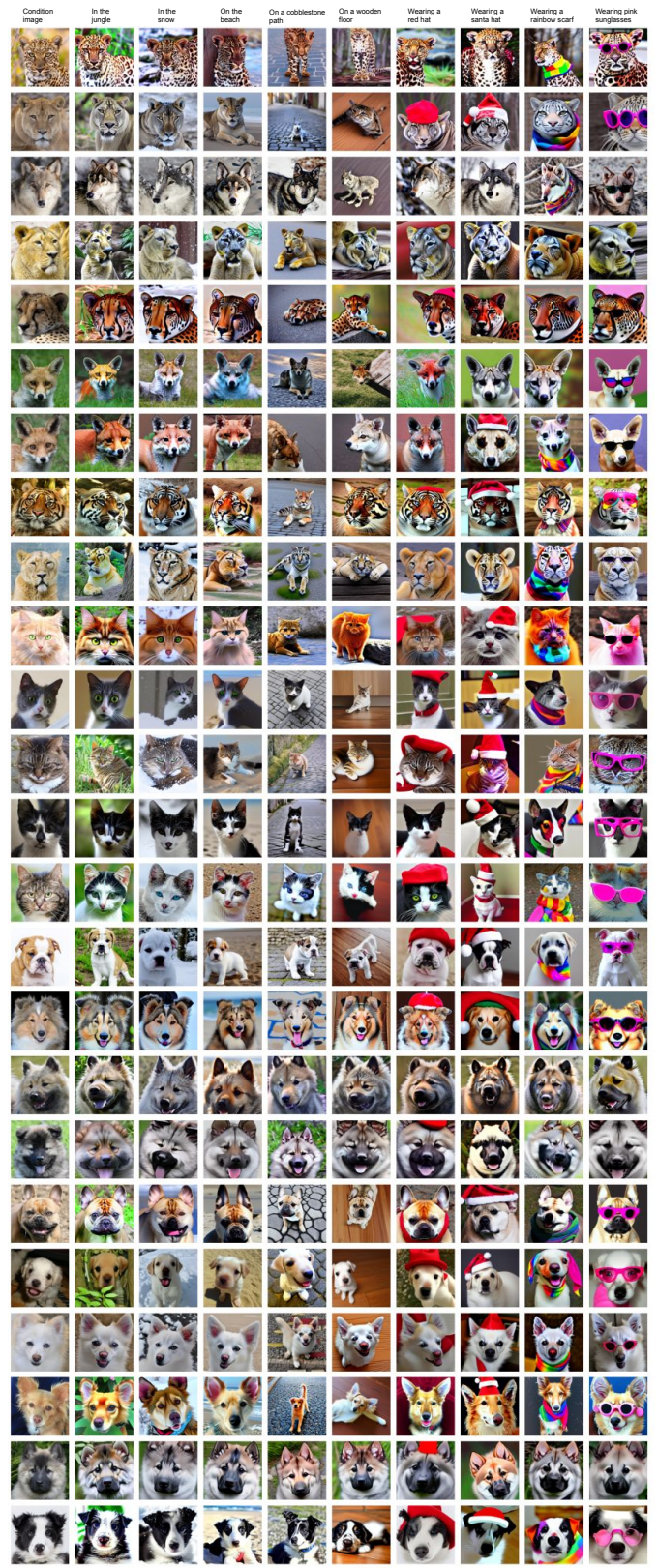


Figure 11. **AFHQ with fast fine tuning** – Each of these images had 50 iterations of DreamBooth fast fine tuning applied to them.



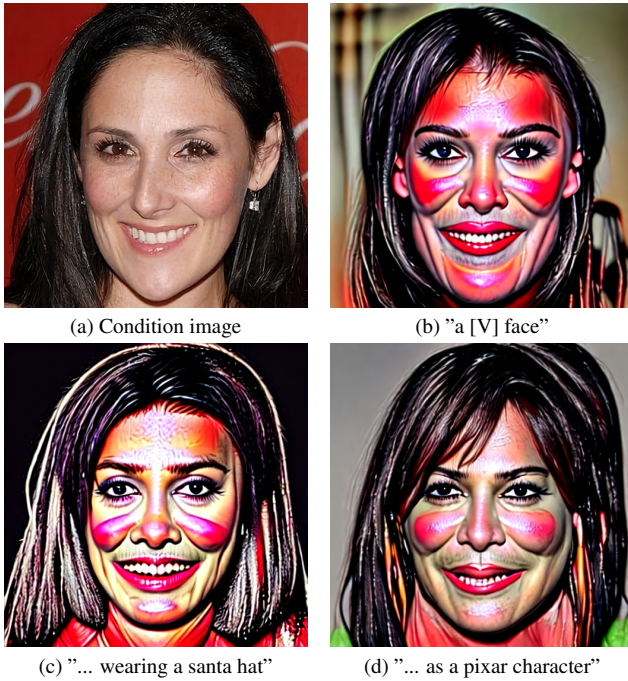


Figure 12. **Example results of training with the task-specific loss** – Training directly with the task-specific loss results in the hypernetworks training simply learning to overfit to a specific training sample, and ignoring the user prompt. As shown, results start being irrelevant to the prompt. Also this further results in unstable training, and thus the results shown with red artifacts on faces.