Supplementary material : Unite and Conquer: Plug & Play Multi-Modal Synthesis using Diffusion Models

Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara and Vishal M. Patel
Johns Hopkins University, Baltimore, MD,USA
{ngopala2, wbandar1 and vpatel36}@jhu.edu

1. Generalized product of experts for differential weighing of different modalities

As mentioned in section the conditional density in the presence of multiple conditioning strategies turns out to be

$$P(z|\mathbf{X}) = \frac{P(z)}{P(\mathbf{X})}\prod_{i=1}^{N} P(x_i|z) = KP(z)\frac{\prod_{i=1}^{N} P(z|x_i)}{\prod_{i=1}^{N} P(z)}, \tag{1}$$

Under Gaussian assumption where the unconditional densities and the conditional densities follow a Gaussian distribution, The above equation could be approximated using Generalized product of experts [1] to

$$P(z|\mathbf{X}) = KP(z)\frac{\prod_{i=1}^{N} q(z|x_i)}{\prod_{i=1}^{N} q(z)}, \tag{2}$$

where $q(.)$ is an estimate of the original density and

$$q \sim N(\mu, \sigma) \tag{3}$$

Generalized product of experts [1] further allows stringer conditions to be applied to the individual conditional densities and reweigh the individual densities to favour some over other, represented by

$$P(z|\mathbf{X}) = KP(z)\frac{\prod_{i=1}^{N} q^{w_i}(z|x_i)}{\prod_{i=1}^{N} q(z)}, \tag{4}$$

Bringing the idea of reliable means in the equation, and assuming that a good estimate of the conditional densities can be made, Eq 4 changes to

$$P(z|\mathbf{X}) = (\prod_{i=1}^{N} P_{\delta_i}^{a_i}(z_t|\phi))\frac{\prod_{i=1}^{N} P_{\delta_i}^{w_i}(z_t|x_i)}{\prod_{i=1}^{N}(\prod_{j=1}^{N} P_{\delta_i}^{a_i}(z_t|\phi))^{w_i}}, \tag{5}$$

Hence the effective score becomes

$$\nabla_{z_t}\log(z_t|\mathbf{X}) = \nabla_{z_t}\log\Bigg((\prod_{i=1}^{N} P_{\delta_i}^{a_i}(z_t|\phi))\frac{\prod_{i=1}^{N} P_{\delta_i}^{w_i}(z_t|x_i)}{\prod_{j=1}^{N}\prod_{i=1}^{N} P_{\delta_i}^{a_i w_i}(z_t|\phi)}\Bigg) = \sum_{i=1}^{N} a_i \nabla_{z_t}\log P_{\delta_j}(z_t|\phi) + \sum_{i=1}^{N}\Bigg(w_i \nabla_{z_t}\log P_{\delta_i}(z_t|x_i) - w_i \sum_{j=1}^{N} a_j \nabla_{z_t}\log P_{\delta_j}(z_t|\phi)\Bigg), \tag{6}$$

The equivalent score can be represented by

$$\epsilon_c = \epsilon_\theta(z_t, \mathbf{X}, t) = \sum_{i=1}^{N} a_i \epsilon_i(z_t, \phi, t) + \sum_{i=1}^{N} w_i\Bigg(\epsilon_i(z_t, x_i, t) - \sum_{j=i}^{N} a_j \epsilon_j(z_t, \phi, t)\Bigg), \tag{7}$$

2. Extension to cases with different variance schedules

Regardless of the variance schedules, one key property in diffusion models is that adding noise equivalent to the T timesteps should always converge to a standard normal distribution, i.e

$$q(x_T|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathcal{I}). \tag{8}$$

where $\bar{\alpha}$ is the cumulative product of the variance schedule β. Once the effective ϵ at each timestep is obtained the individual scores could be calculated using the equation for score calculation mentioned in the main paper and the image at the next timestep could be obtained using the the respective variance schedule.

3. A new approach for text to image generation for facial images

In all of our experiments defined in section we include text based description one of the modalities. The existing diffusion based approaches make use of millions of paired image-text pairs and train models for the task of text to Image generation. But in multimodal scenarios because of the availability of information from other modalities, we do not require such a powerful model. Hence we propose an novel lightweight text to image approach as follows. Consider a conditional diffusion model for text to image conversion defined by $g_\phi(z_t, \text{emb}, t)$. For our model we make use of a image-text pretraining model like CLIP with the corresponding Image and text embedders defined by $(h_{img}(.), h_{text}(.))$. During training, we condition using the low dimensional Image emdedding using adaptive group normalization at stage of our network. The Image embedding is obtained by using the image embedder of clip on the incoming training image. The loss for optimization of the parameters ϕ of the network are defined by,

$$L = E_{t\sim[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - g_\phi\left(\mathbf{z}_t, h_{img}(z_0), t\right)\|^2\right] \quad (9)$$

During the inference process, we make use of the text embedder of the vision-language model and obtain the text embeddings by passing the corresponding text conditioning $text_{in}$ sample using the process,

$$z_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} g_\phi(z_t, h_{text}(text_{in}), t)\right) + \sigma_t^2 \boldsymbol{\eta},$$
$$\text{where } \boldsymbol{\eta} = \mathcal{N}(\mathbf{0}, \boldsymbol{I}),$$

The embeddings for image and text were obtained by using [4]

4. Text to Image generation:-

For Multimodal text to Image generation, we comparing with existing works trained for text to image generation in CelebA-MM dataset. But unlike the exiting works that were trained with paired text-image prompts, our model perform zero-shot text to image generation and it has never seen any text prompts during training time. We utilize FARL [4] for obtaining the image-text embeddings of the images. For evaluating the quality of text to image generation, we follow [3].

5. Sketch to Face training:-

For generating rough sketches of the faces, we utilize [2] edge detector to extract edges from the image and use this as the conditioning technique while training the network.

References

[1] Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014. 1

[2] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 2

[3] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 2

[4] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109*, 2021. 2

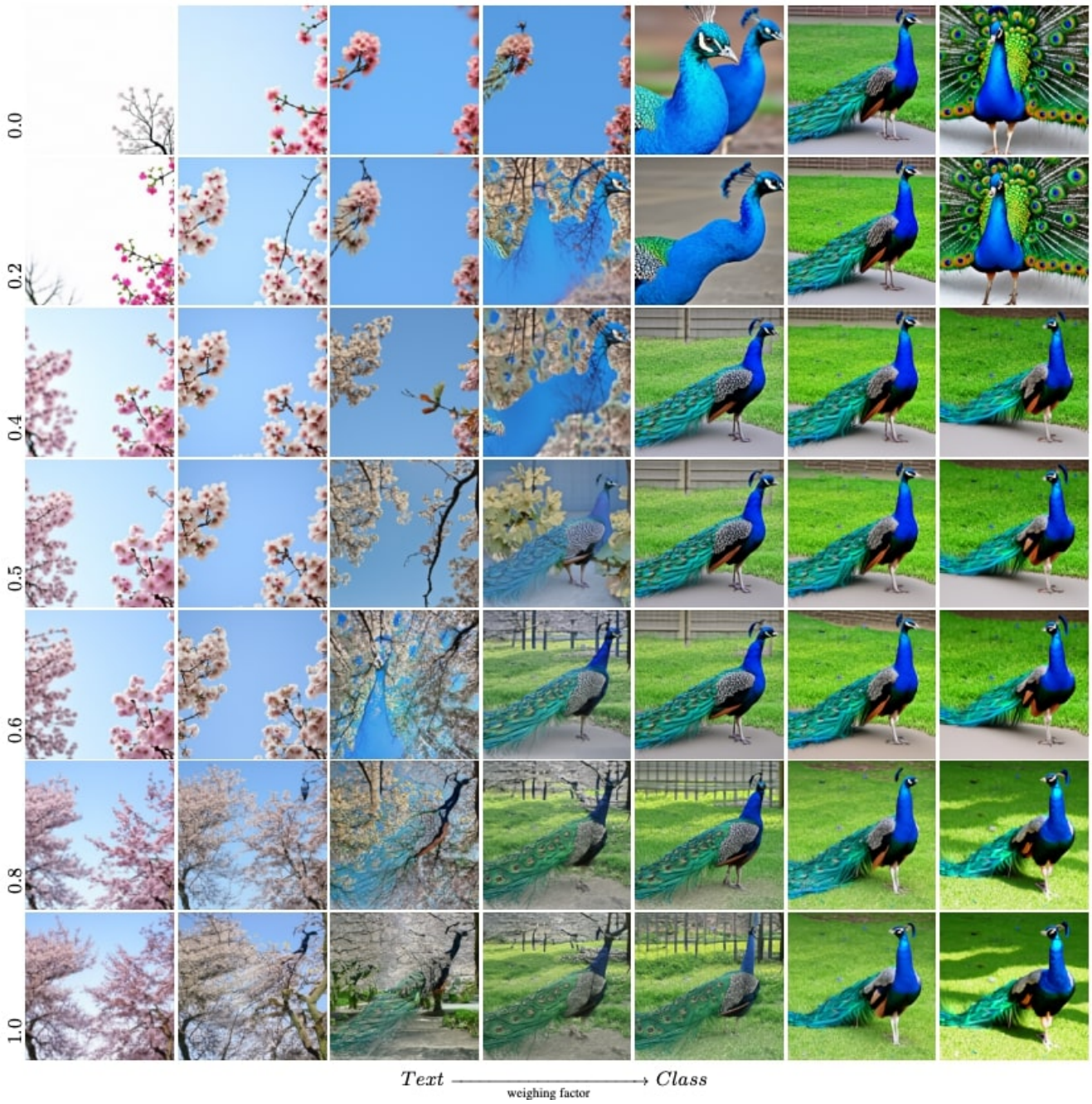

Figure 1. **The role of reliability factor for multimodal generation.** Text Prompt is ”A garden of cherry blossom trees, Class is ”84: Peacock”. y-axis, higher the value, higher the reliability for the text model

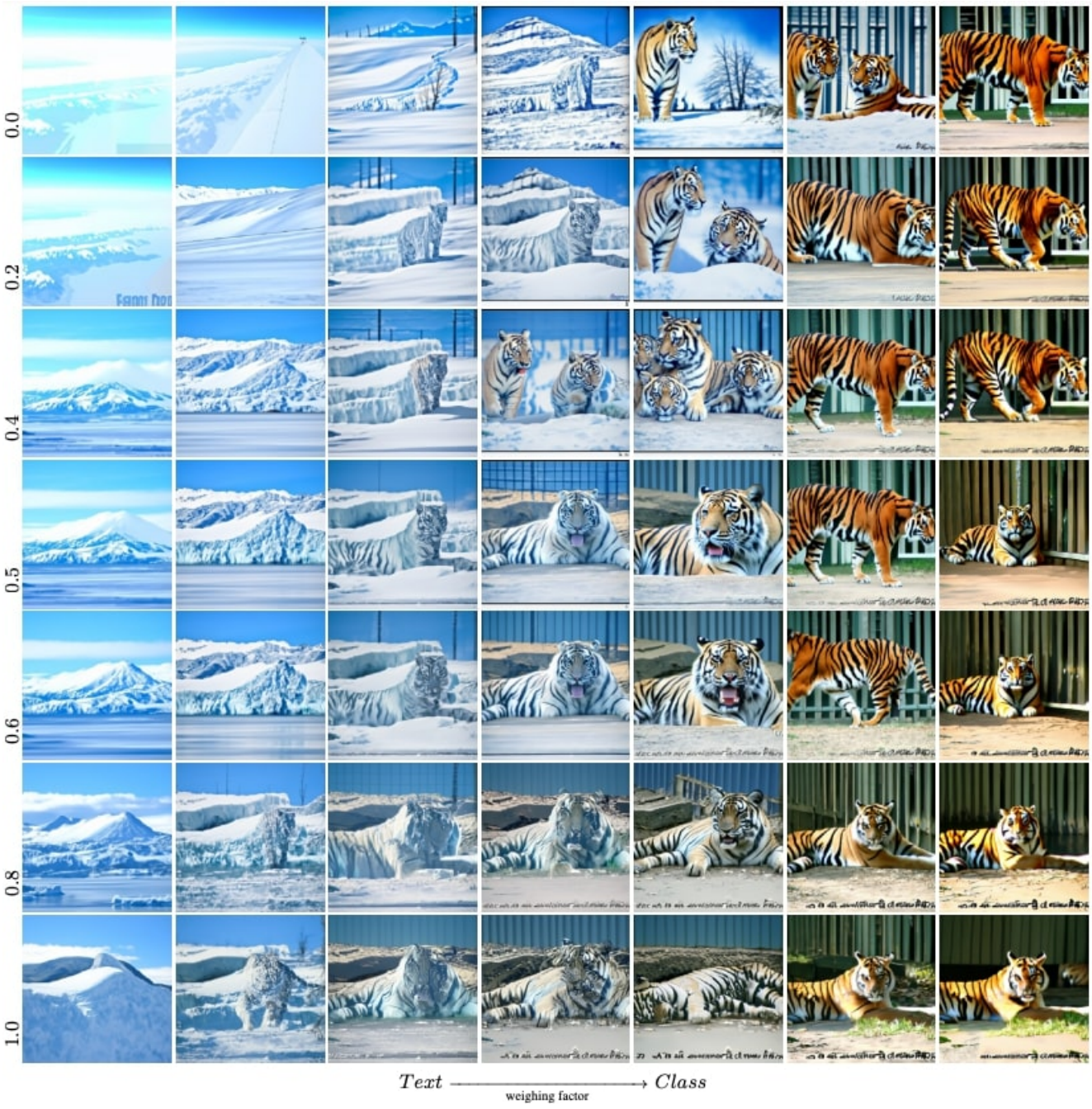

Figure 2. **The role of reliability factor for multimodal generation.** Text Prompt is ”A ice mountain, Class is ”292: Tiger” y-axis, higher the value, higher the reliability for the text model

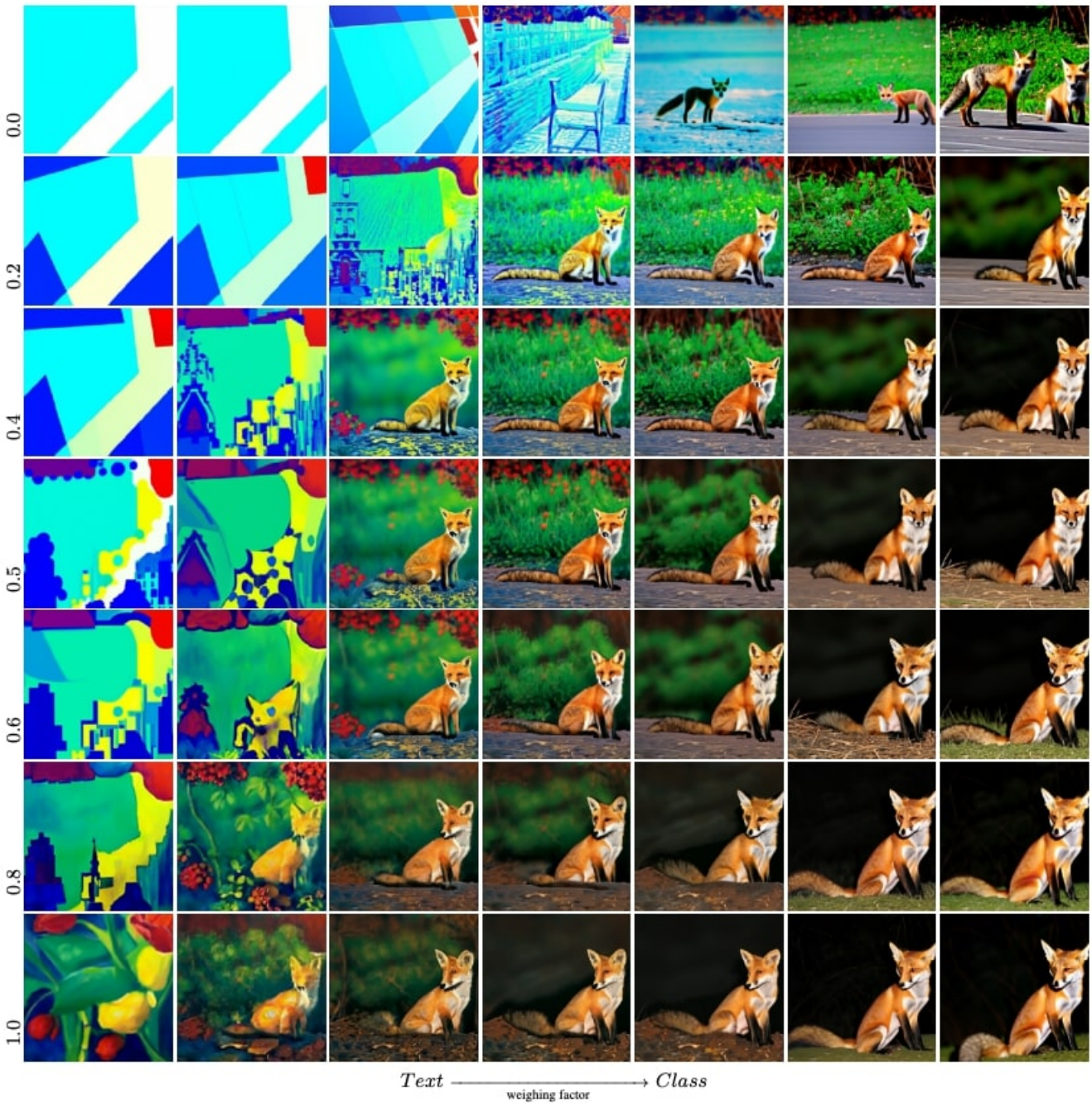

Figure 3. **The role of reliability factor for multimodal generation.** Text Prompt is ”An oil painting, Class is ”277: Red Fox”. y-axis, higher the value, higher the reliability for the text model

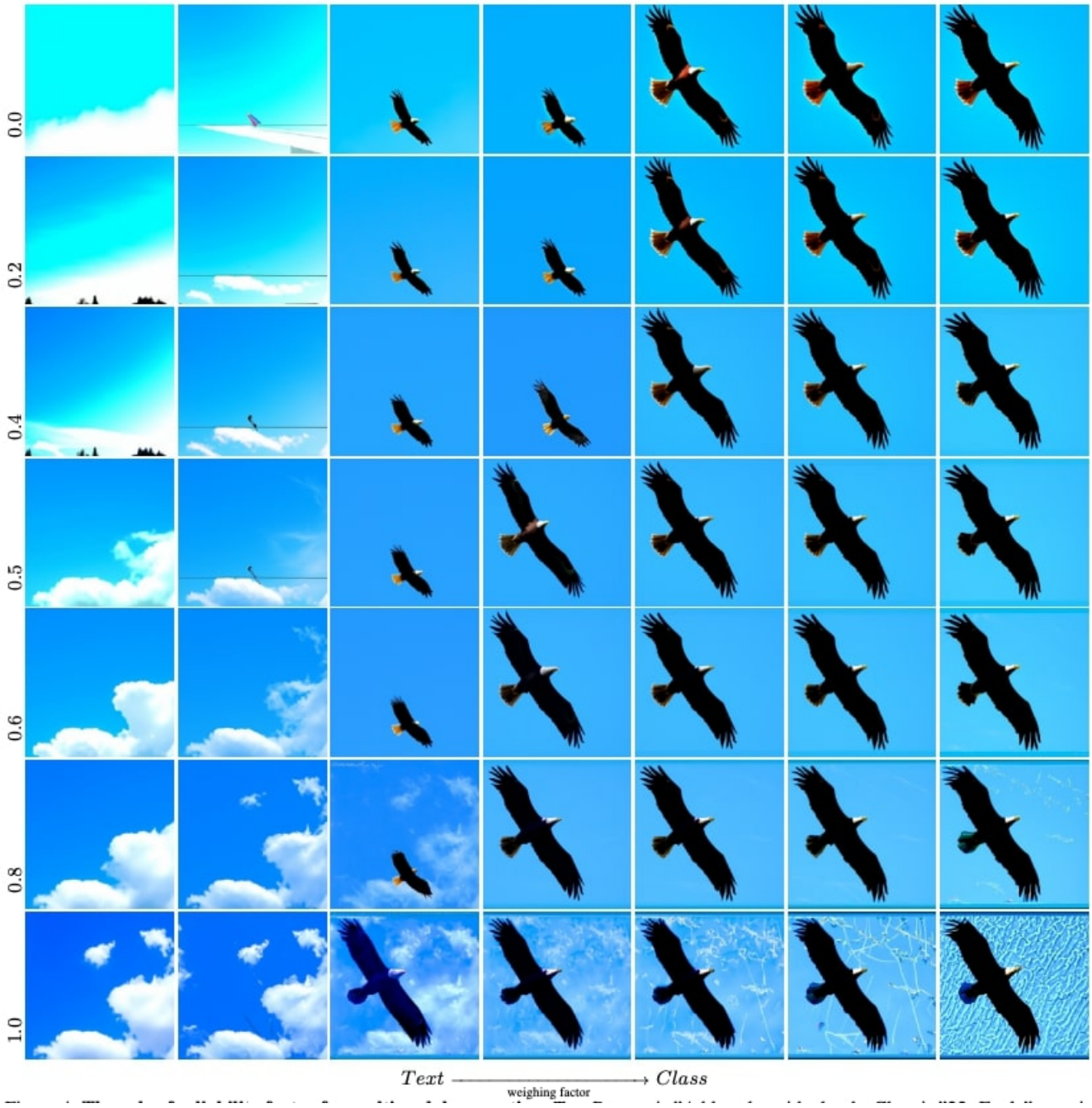

Figure 4. **The role of reliability factor for multimodal generation.** Text Prompt is ”A blue sky with clouds, Class is ”22: Eagle”. y-axis, higher the value, higher the reliability for the text model

Figure 5. Results for text ”A crayon drawing” and ImageNet class ”663: Monastry”.Not cherry-picked

Figure 6. Results for text ”A snow mountain” and ImageNet class ”277:- Red fox”.Not cherry-picked

Figure 7. Results for text ”A cherry blossom tree” and ImageNet class ”17: Jay”. Not cherry-picked

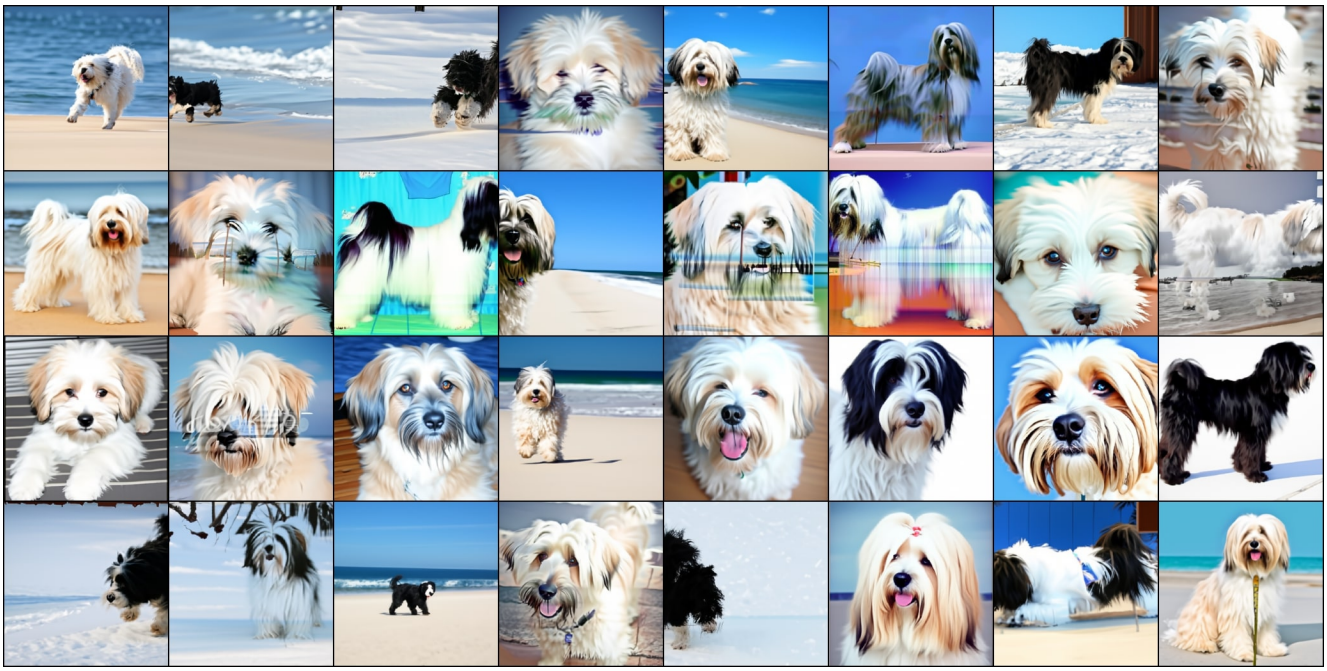

Figure 8. Results for text ”Photo of a beach” and ImageNet class ”200:- Tibetian Terrier”.Not cherry-picked

Figure 9. Results for text ”A wheat field” and ImageNet class ”291:- lion”.Not cherry-picked

Figure 10. Results for text ”A road leading into mountains” and ImageNet class ”483: A castle”. Not cherry-picked

Figure 11. Results for text ”Photo of a beach” and ImageNet class ”385: Elephant”. Not cherry-picked

Figure 12. Results for text ”A canoe on the sea” and ImageNet class ”442: A bell cote”. Not cherry-picked

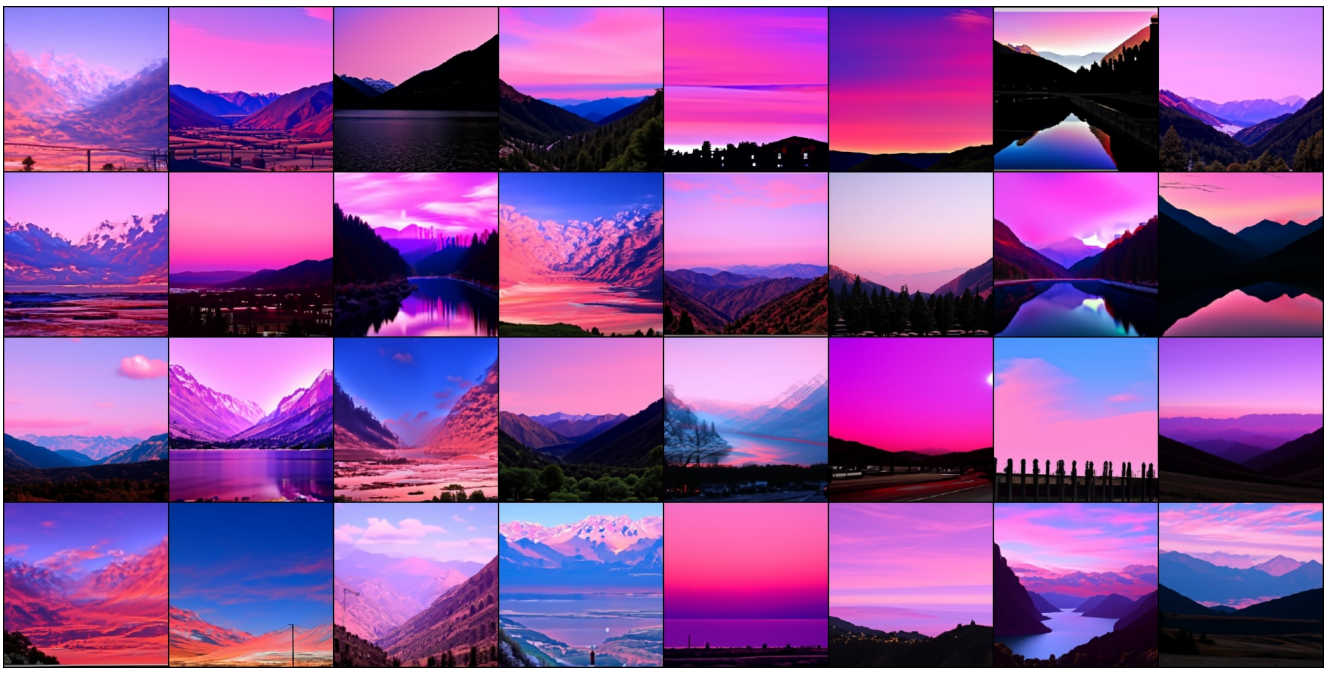

Figure 13. Results for text ”A pink sjy” and ImageNet class ”979: A valley”. Not cherry-picked

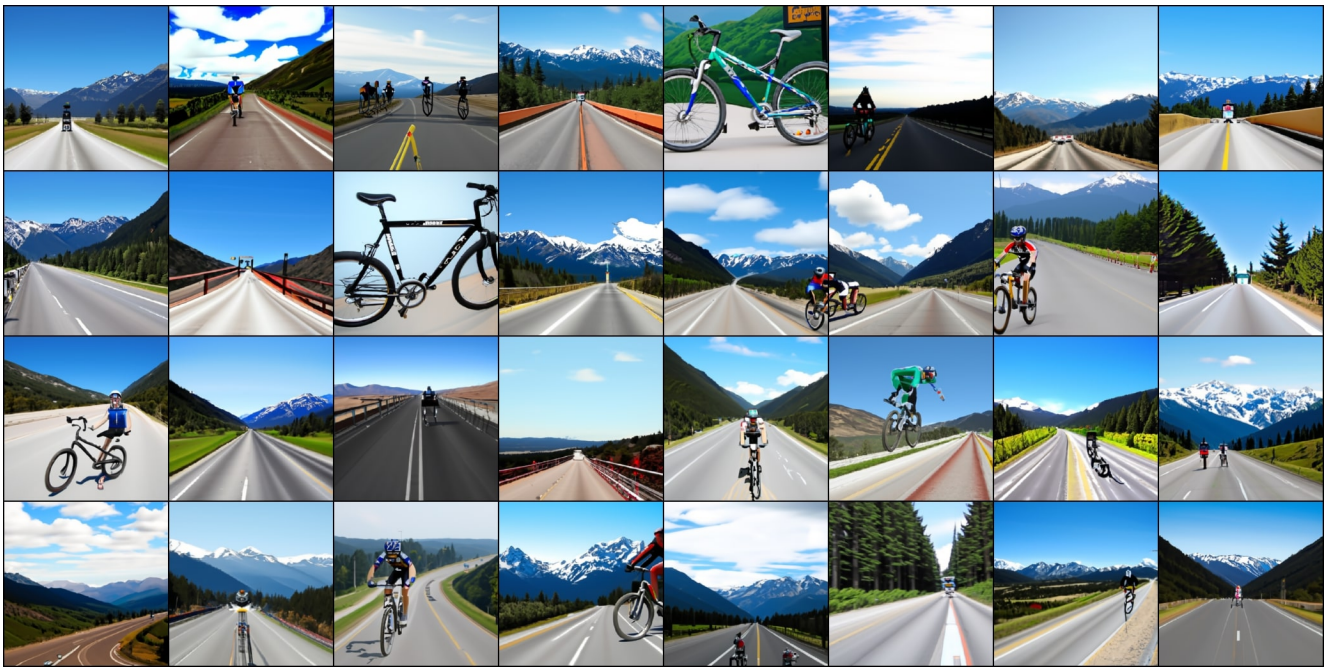

Figure 14. Results for text ”A highway” and ImageNet class ”671: A bike”. Not cherry-picked

Figure 15. Results for text ”A road leading to mountains” and ImageNet class ”850: Teddy bear”. Not cherry-picked

Figure 16. Results for text ”A paddy field” and ImageNet class ”345: ox”. Not cherry-picked

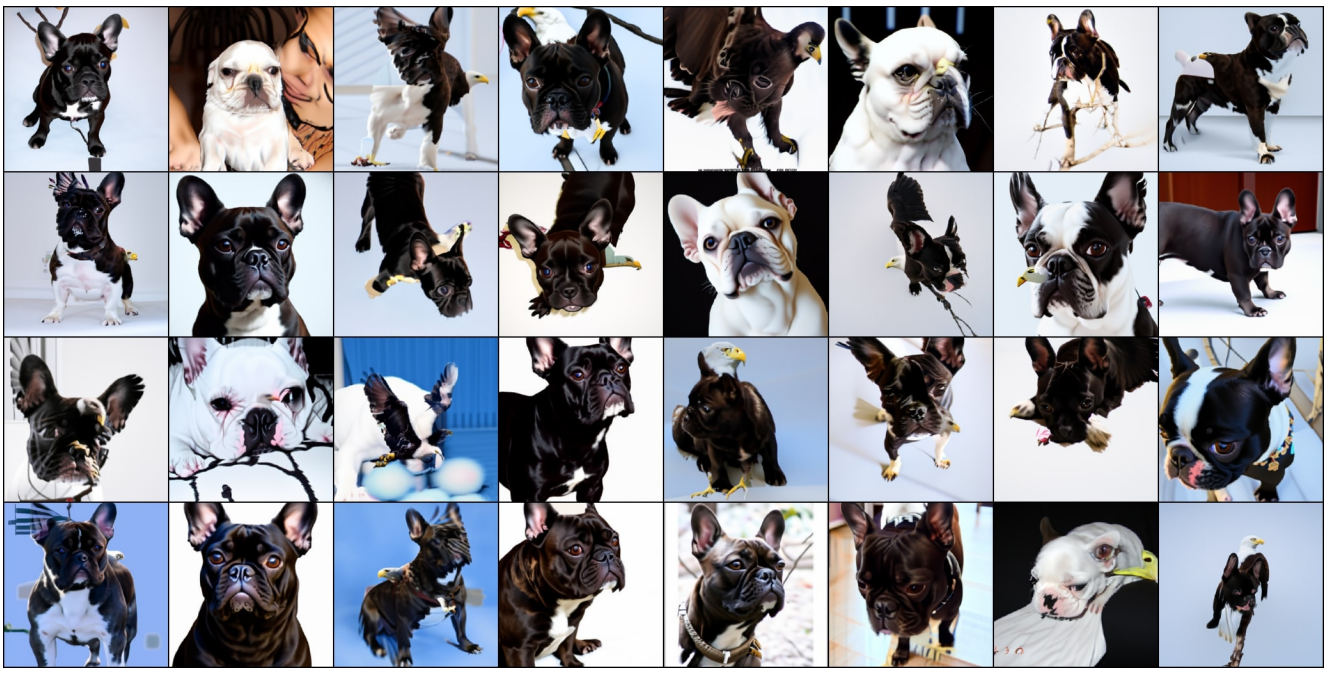

Figure 17. Failure case: Contradictory inputs ”An Eagle” and ImageNet class ”245: Bulldog”.

Figure 18. Failure case: Contradictory input ”Photo of a cat” and ImageNet class ”96: Toucan”.

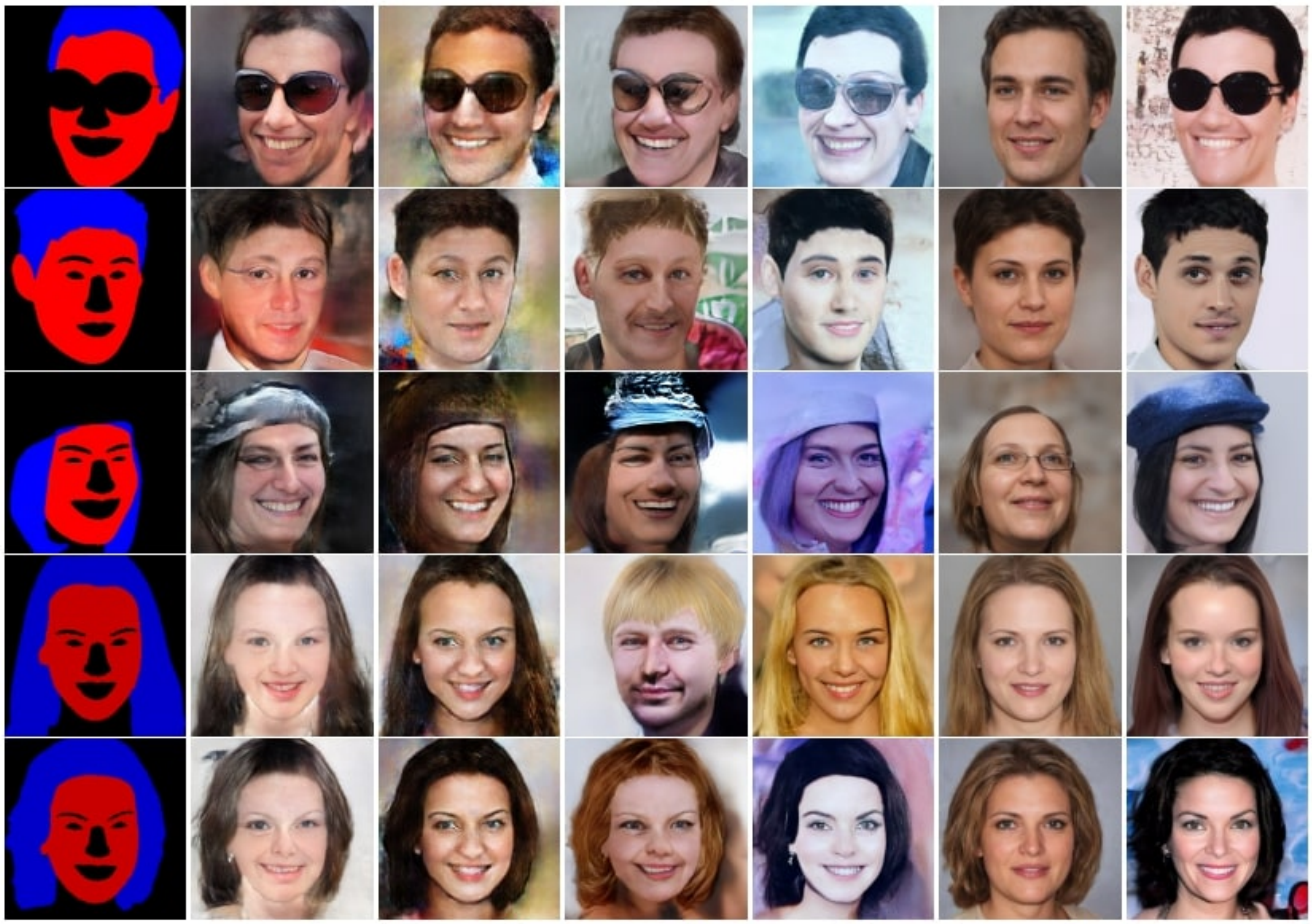

Figure 19. **Qualitative comparisons for semantic to face generation.** In this case, a single model is trained by alternating different input datasets across different iterations. During Inference time all the modalities are taken from a single dataset and the proposed sampling technique is used.

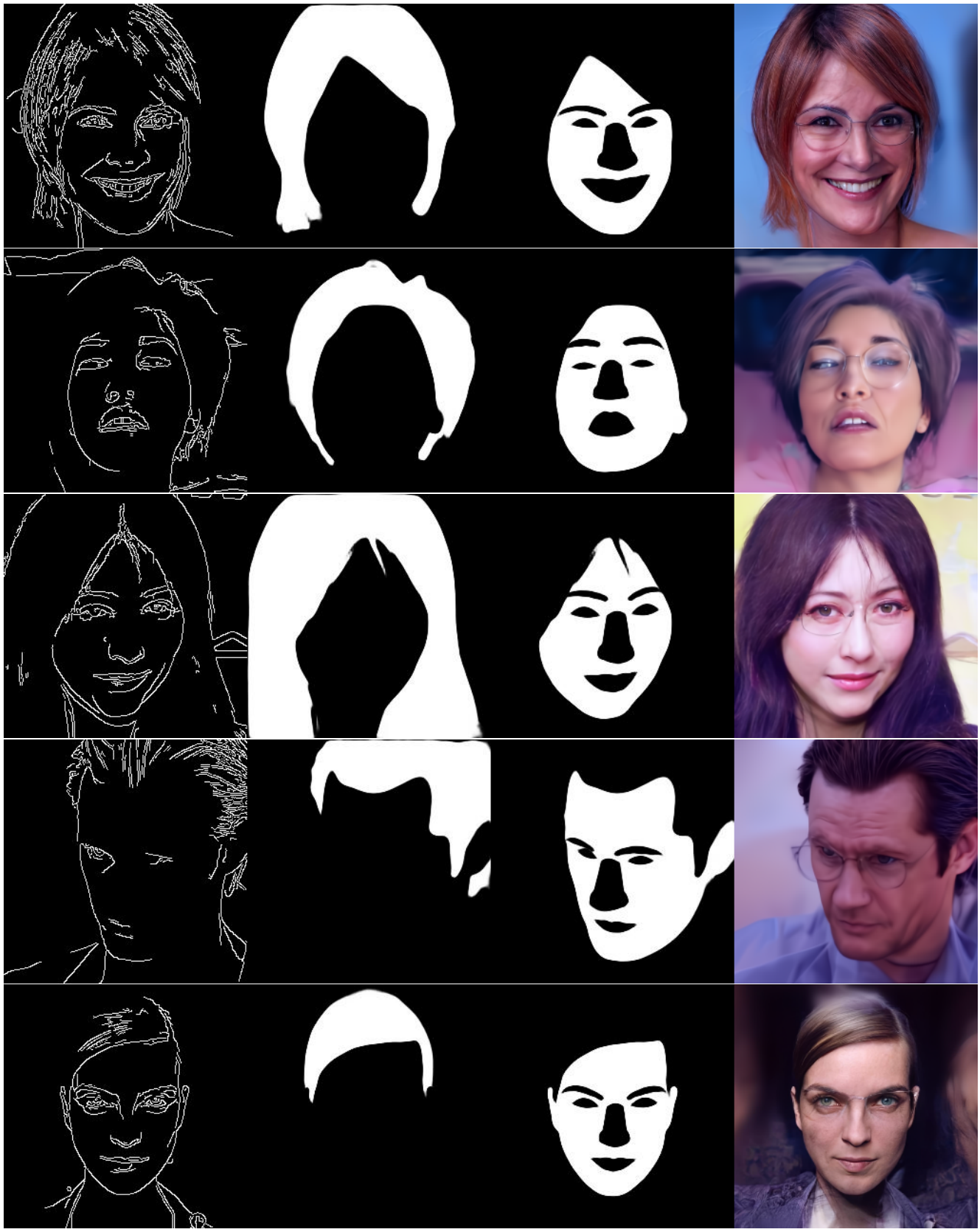

Figure 20. **Multimodal face generation using four modalities** Text used: "An old person with eyeglasses"

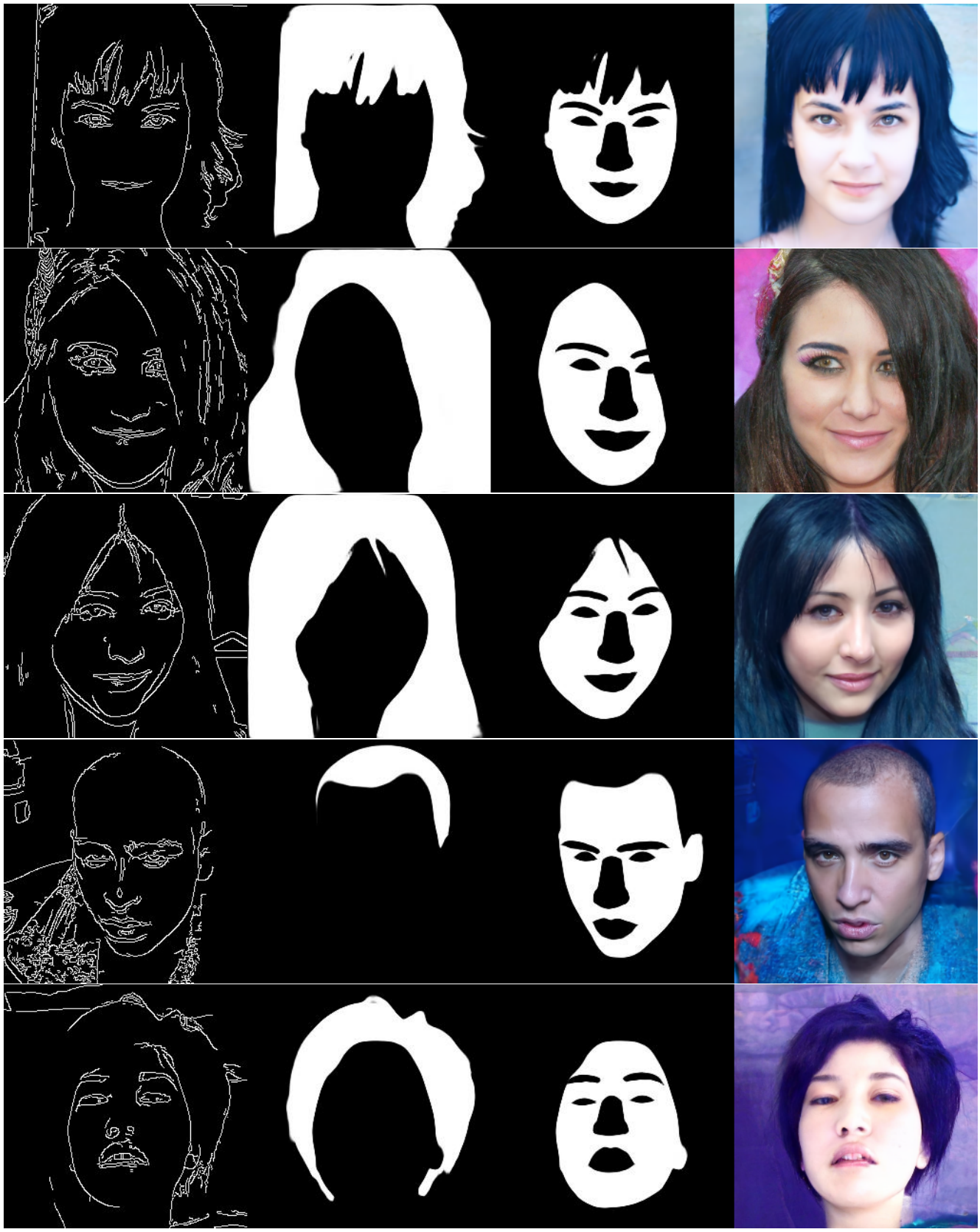

Figure 21. **Multimodal face generation using four modalities** Text used: "A person with black hair"

Figure 22. **Multimodal face generation using four modalities** Text used: "A person with eyeglasses"

Figure 23. **Multimodal face generation using four modalities** Text used: "A person with blonde hair"

Figure 24. **Multimodal face generation using four modalities** Text used: "This person has gray hair and wears eyeglasses"

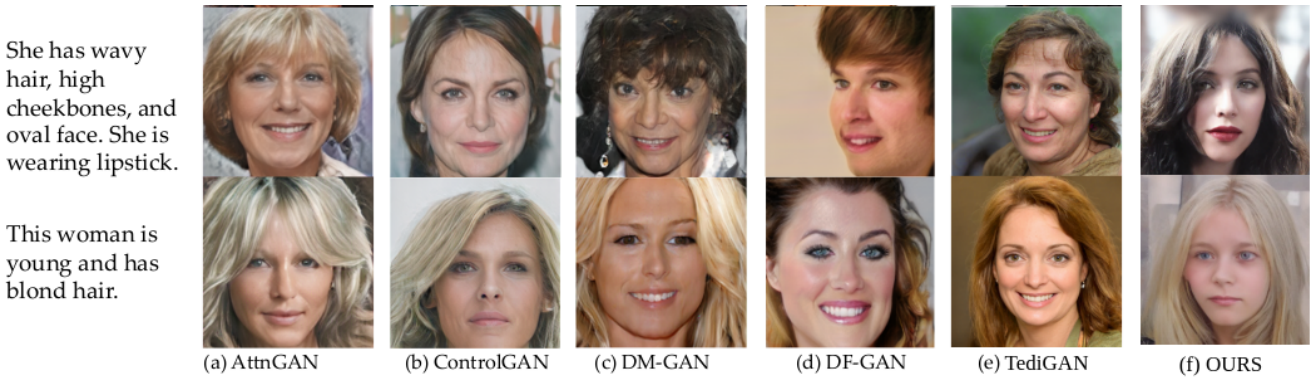

Figure 25. Results for text to image generation on CelebA-HQ dataset.