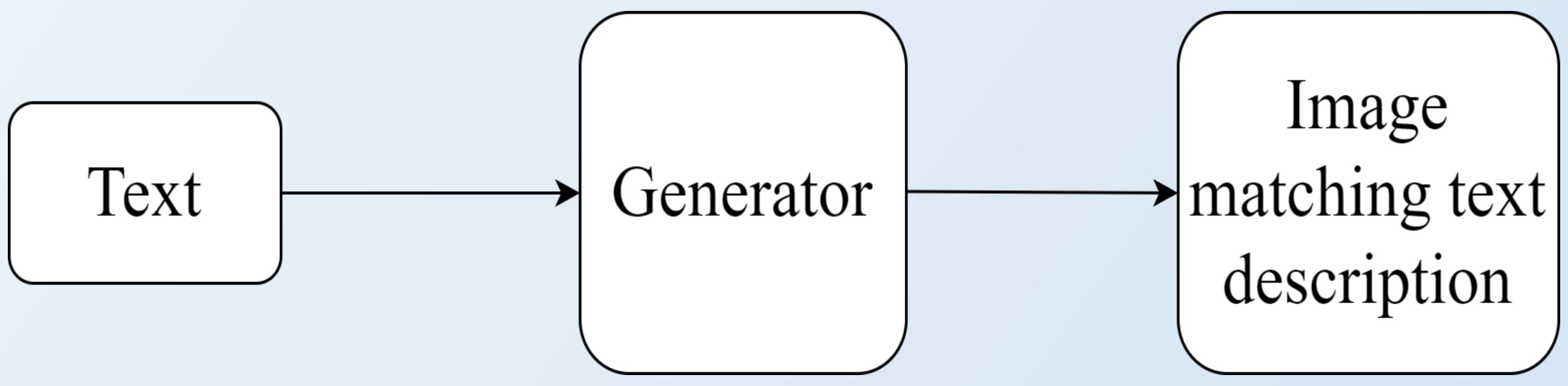
Adversarial Text-to-Image Synthesis

By Ahmed Tammaa, Supervised by: Assoc. Prof. Nahla Barakat

Faculty of informatics and Computer science, Artificial Intelligence Department, The British University in Egypt

Abstract

Text-to-Image generation is an open research field that aims to generate a photo-realistic image from a text description. It demonstrates the power of computer imaging and image editing. We showed that using a large embedding vector can improve the image quality at the expense of diversity. This project studies an experiment of instance selection preprocessing that made the model trains faster. However, the performance has degraded due to the tradeoff between speed and quality. We demonstrated why and how hierarchical structures improve the quality of the generated image. We have achieved the best **FID of 258 and Inception score of 2.43 using attention GAN** with fewer epochs than the original.



Summary of the project idea

Methodology

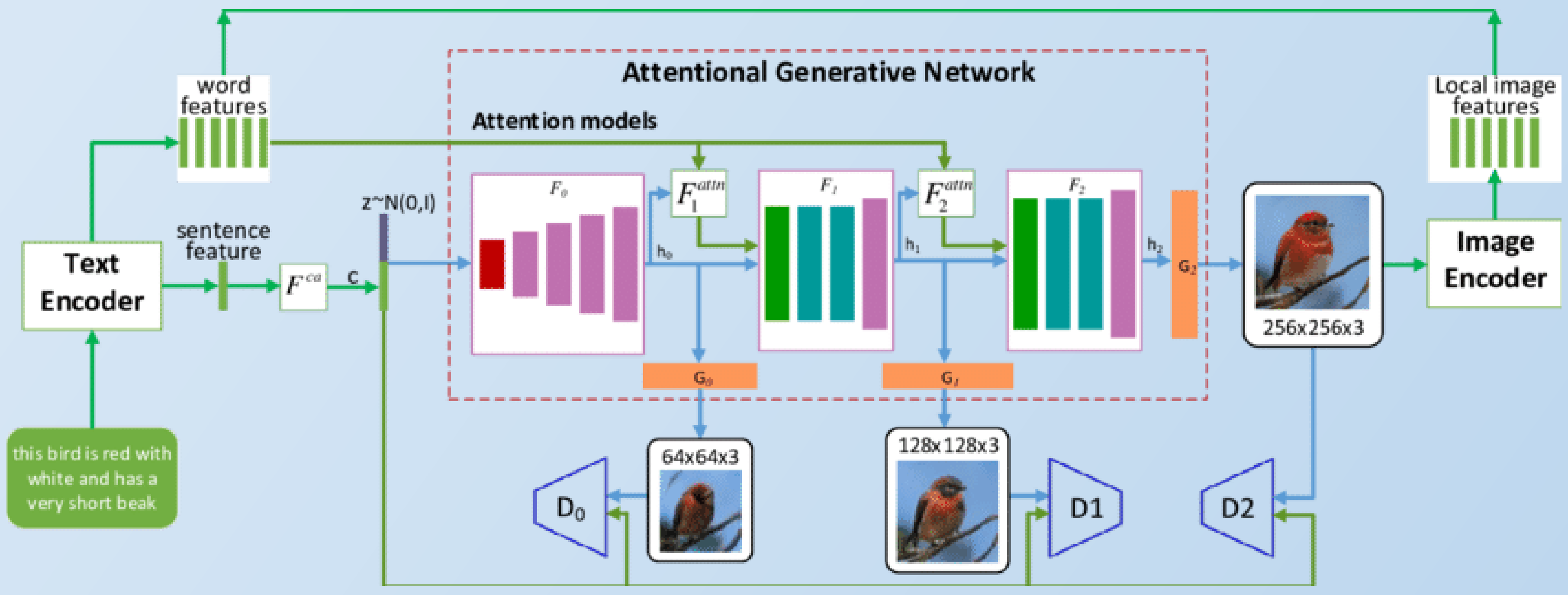
Method 1: Use Matching-Aware Discriminator (GAN-CLS) and Interpolate data manifolds by the generator (GAN-INT). Combining both architectures (GAN-INT-CLS)

Matching-Aware Discriminator: Naïve GAN discriminator can only know if the image is real or fake. Once the generator learns to generate realistic images, the discriminator starts to accept them even if the text does not match. GAN-CLS modified the algorithm to refuse the image if it does not match the text description. The training operates as shown in the table below.

Image statues	Caption Statues	Correct Prediction
Fake Image	Mismatching caption	**Fake**
Real Image	Mismatched Caption	**Fake**
Fake Image	Matching Caption	**Fake**
Real Image	Matching Caption	**Real**

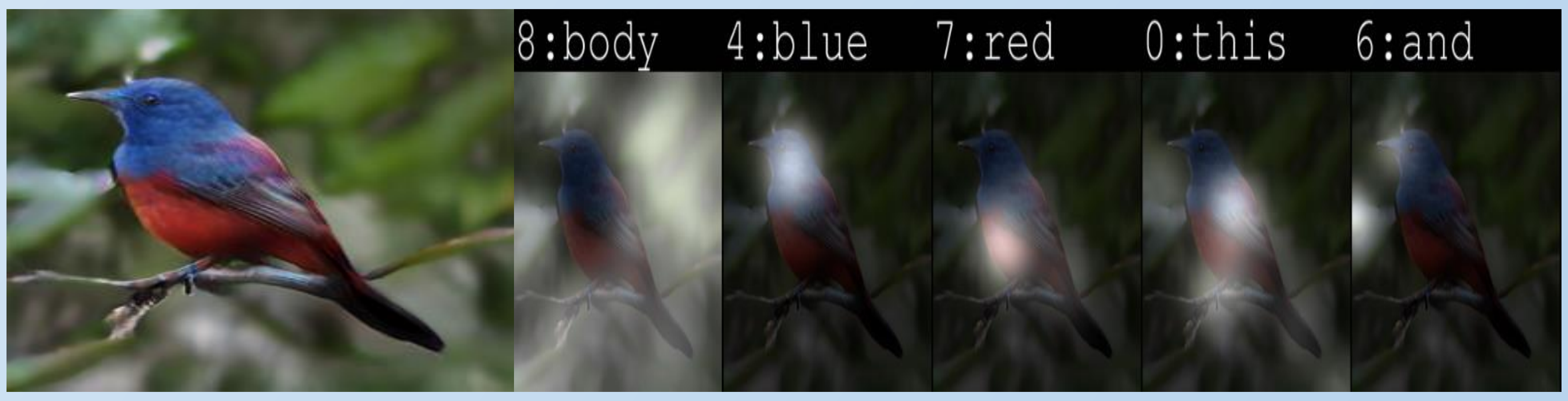
Interpolation: Generate additional captions by interpolating between two embeddings even if they came from different categories. Subsequently, the discriminator does not have a "real" image corresponding to the interpolated captions. It should inspect if the generated image matches the text input. If the discriminator did well in this task, the generator learns to fill the gaps in the data manifolds.

Method 2: Use attention to measure the Multimodal similarity between the text and the image. This approach makes the generator draw sub-regions based on the text description, such as "Blue head." which will assure the generator goes to the "head" region and paint it blue. This architecture is hierarchal. Therefore, every step is dependent on its previous. The first image contributes to the final image.



Attention GAN architecture

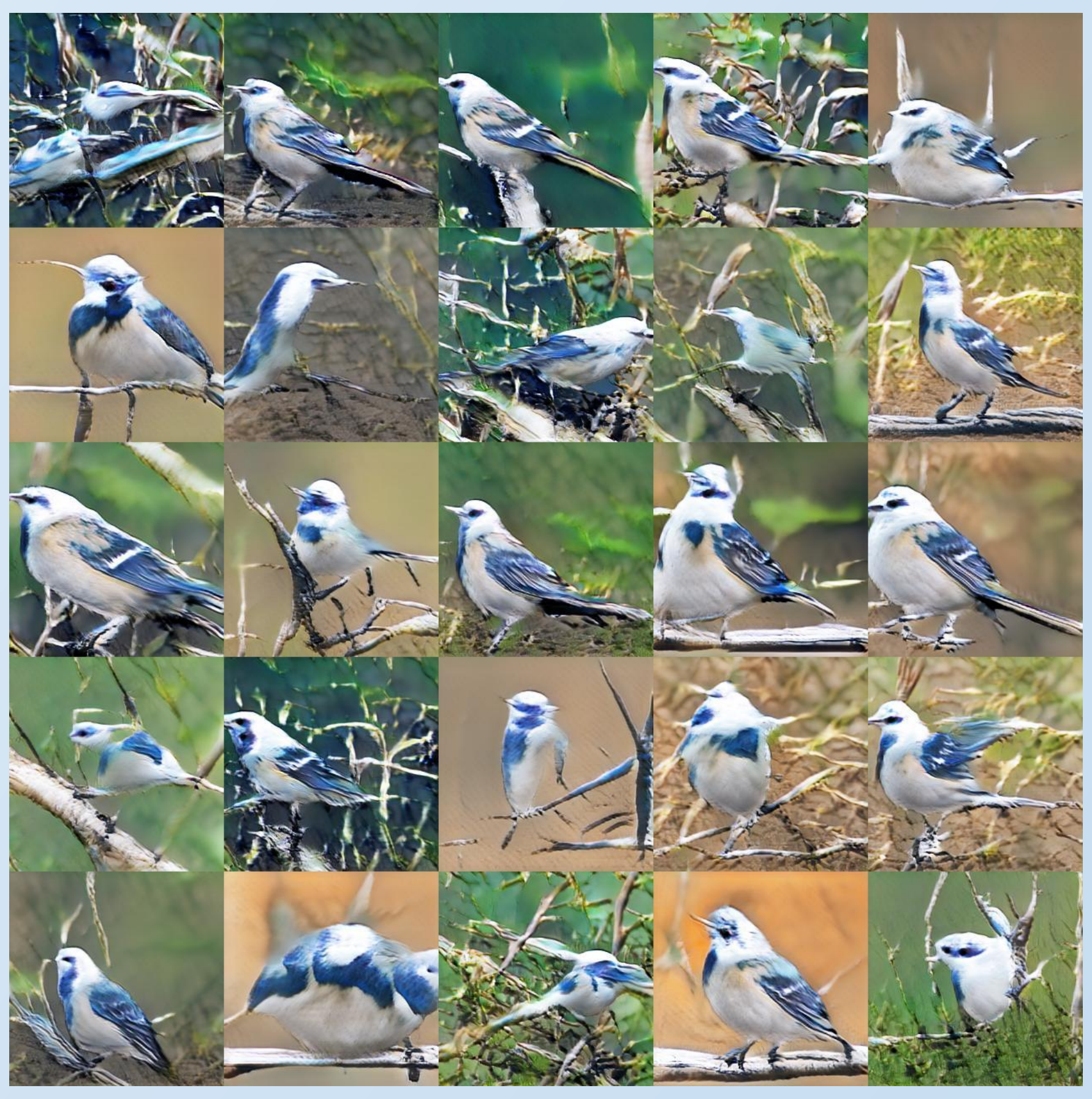
Below illustrates how the attention model cares for the caption: "this bird has a blue head and red body"



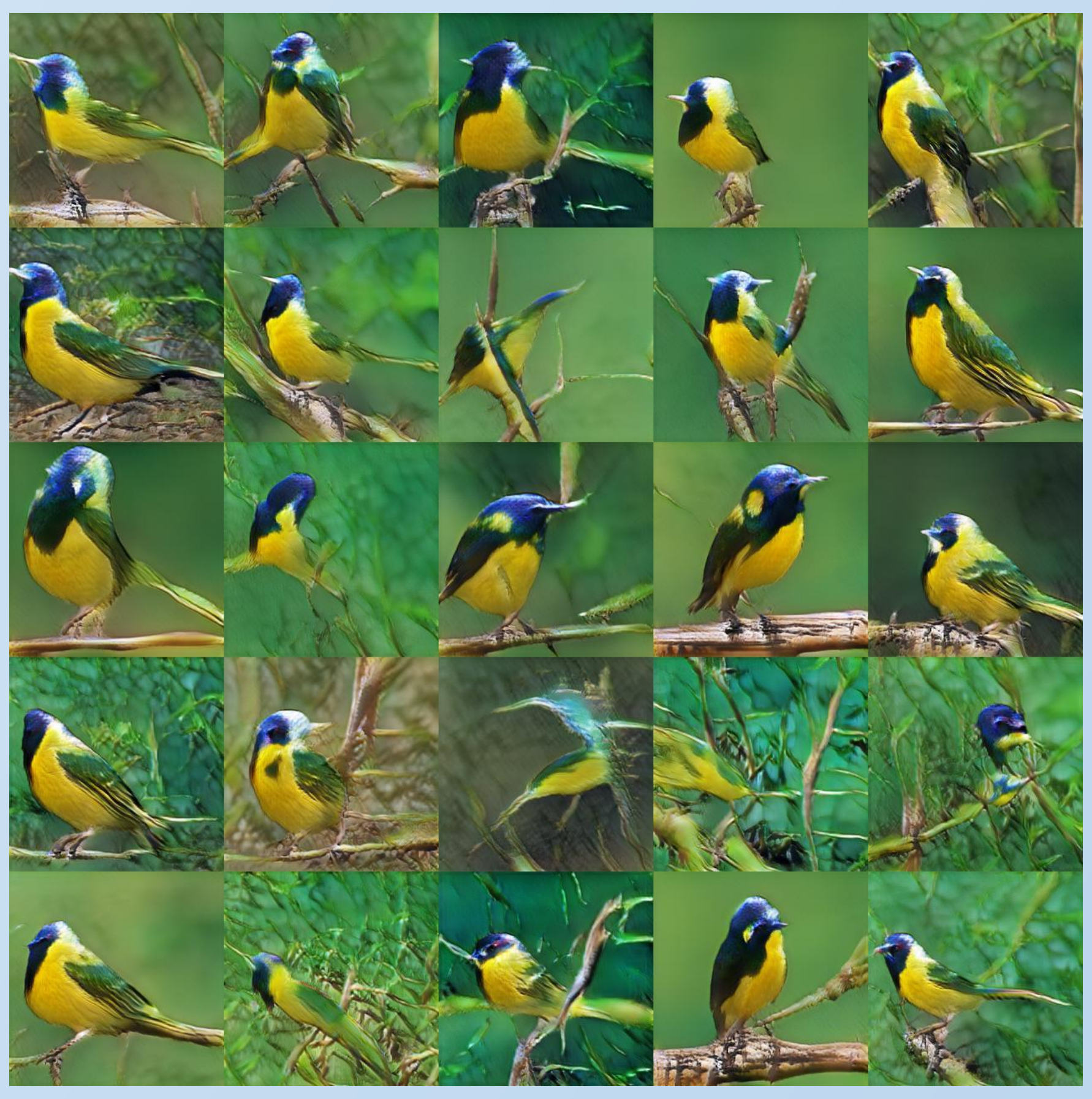
Attention plot for the prompt "this bird has a blue head and red body

Results

Qualitative Results: Below is two examples of generated images. "This bird has a white head and blue body" (left), "This bird is yellow with a blue head" (right).

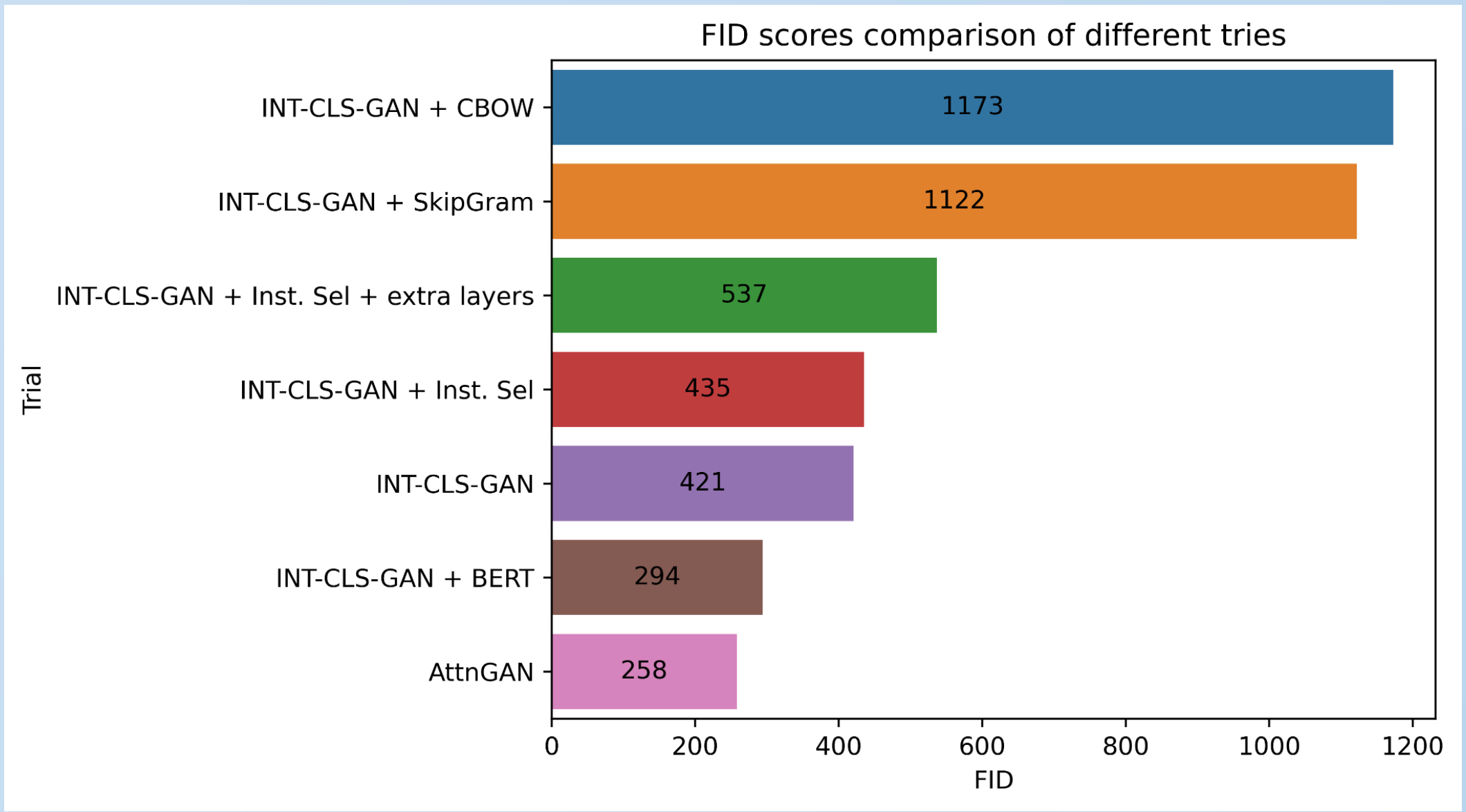


Output for the prompt "This bird has a white head and blue body"



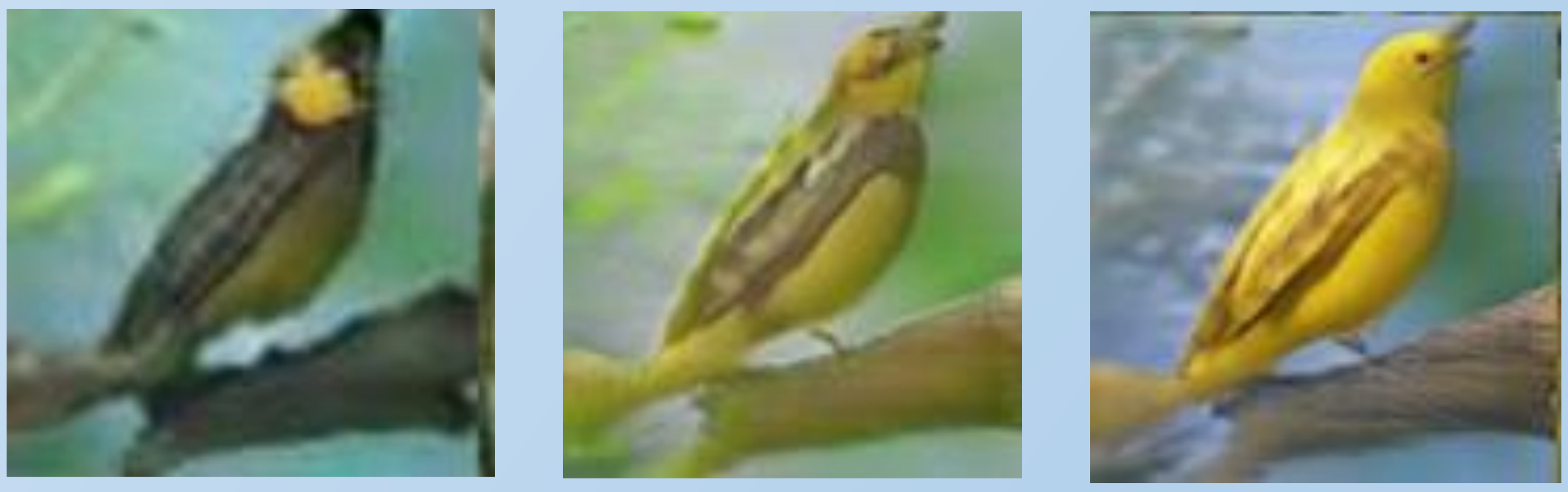
Output for the prompt "This bird is yellow with a blue head"

Quantitative Results: Below is a comparison between different trials on the FID metric in the project (the lower the value, the better the score)



FID scores for project's different trials

Figure below elaborate the hierarchal output of our AttnGAN. The images were upscaled to the same size for clear comparison. Scales from left to right 64 × 64 → 128 × 128 → 256 × 256



Generation output of "This bird is yellow" from 64 × 64 to 256 × 256

Conclusion

- Generating a 64×64 image is the critical step of image fidelity.
- Hierarchical Structure of GAN outperforms deep architectures
- Instance Selection helps for both downsizing and allows for more complex structures.
- Creating Word2Vec embedding for the CUB birds narrows down the diversity and fidelity of the images.
- Usage of higher dimensional embedding vector reduces the text-image semantic alignment; however, it scored higher FID.
- Improving the discriminator ability improves the generator's ability to be more intelligent than generating a realistic image.
- Attention proves to be essential for improving the text-image semantic alignment and has the best FID score of 258.

References

[1] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.

[2] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).