# Imbalance in Balance: Online Concept Balancing in Generation Models

## Supplementary Material

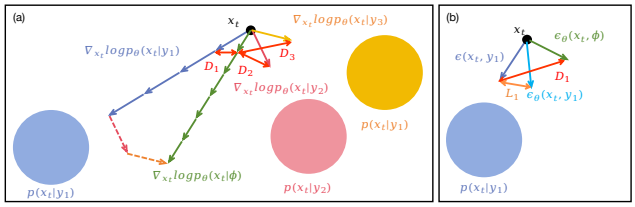


Figure 9. **Formulation of IMBA distance.** (a) Unconditional distribution shifts toward head concepts due to data imbalance, leading to a smaller IMBA distance $D$ for head concepts. (b) Relationships between IMBA distance and diffusion loss during training.

## A. Formulation of IMBA Distance

Based on the above analysis, we can formulate the IMBA distance under the imbalanced data during training in Figure 9. As shown in Figure (a), starting from the random noise $x_t$ in the latent space, conditional distribution points to different data distributions with different color based on different concepts. Due to data imbalance, concept $y_1$ has far more samples than other concepts. Since the unconditional distribution is weighted by all samples equally during training, it will shift toward concepts with more samples like the green arrow, leading to a smaller IMBA distance $D_1$. In the training set, the ratio of samples between head and tail concepts often reaches a factor of thousands, far exceeding the ratio shown in the figure, indicating a much more severe data imbalance issue and more pronounced pattern of IMBA distance. As shown in Figure (b), original diffusion loss represents the distance between the conditional distribution and the predicted conditional distribution, and IMBA distance represents the distance between the predicted conditional distribution and the unconditional distribution. Specifically, when IMBA distance is implemented with the L2 norm, it is equivalent to the unconditional loss.

## B. Ablation study

### B.1. Stability of IMBA Distance

In Figure 10, we calculate the IMBA distance of the same prompt on models with different size, architecture and noise. We find it is stable across all settings.

### B.2. Comparison with Frequecy-based Method

Since the text is a joint distribution of multiple concepts, it is difficult to calculate weights from a frequency perspective, and there is little concept-balancing work for text-to-image generation. Therefore, we compare IMBA loss with a frequency based method on class-image [9]. We sample 5 concepts each from the head and tail concepts and combine the data containing these concepts in the training set into a new subset. We then finetune the model on the subset using the frequency-based and our method respectively. Meanwhile, we pair the 10 concepts to generate 5 captions for each pair as the test set. As shown in the Table 4, our method outperforms the frequency-based method.

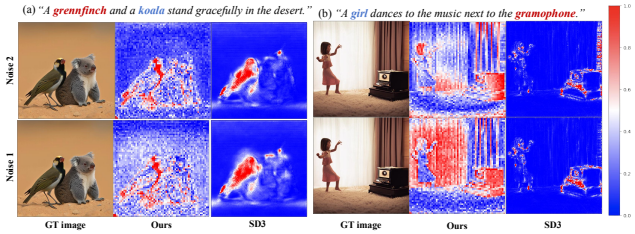


Figure 10. **IMBA distance of different models and noises.**

| Loss weight | Baseline | Frequency-based | Ours |
|---|---|---|---|
| Success rate | 33.3% | 49.3% | 65.7% |
| CLIP Score | 0.3113 | 0.3101 | 0.3218 |

Table 4. The performance of different balancing methods.

| Loss weight | Baseline | Sample-wise | Token-wise |
|---|---|---|---|
| Success rate | 32% | 64% | 72% |
| CLIP Score | 0.2924 | 0.3022 | 0.3106 |

Table 5. The performance of models with different loss weight.


Figure 11. The performance of models with different loss weight.

### B.3. Comparison with Sample-wise Loss Weight

We finetune the same model on the imbalanced "piano-submarine" subset for 10K steps with sample-wise and our token-wise loss weight respectively. As shown in Table 5 and Figure 11, all results are evaluated on 25 captions. And sample-wise loss weight performs better than the baseline due to the reweight balancing. Meanwhile, token-wise loss weight achieves the best performance since it applies more fine-grained weights on different image regions according to concepts.

### B.4. Hyper-parameters of IMBA Loss

We train the model on the "piano-submarine" subset to conduct ablation experiments on the value of $\gamma$. Specifically, when $\gamma = 0.0$, IMBA loss is equivalent to the original dif-


Figure 12. **Results from models trained with different $\gamma$.**

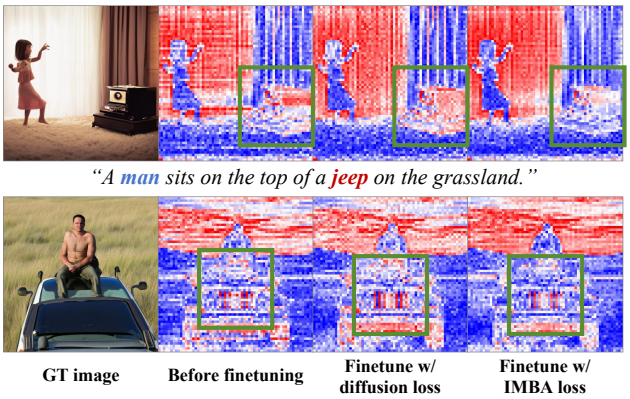


Figure 13. IMBA distance before and after finetuning.

fusion loss. When $\gamma = 2.0$, the value of the IMBA distance equals the value of the unconditional loss. As shown in Figure 12, when $\gamma$ approaches 0.0, the concept composition ability of the model diminishes, as the semantic of the submarine in Figure(a) almost disappears. When $\gamma$ approaches 2.0, the model exhibits severe color shift issues as seen in Figure(b). We chose $\gamma = 0.8$ based on these observations.

### B.5. IMBA Distance after Training.

We resumed training a model for 3 epochs using diffusion loss, and then fine-tuned it separately with diffusion loss and IMBA loss. The difference in IMBA distance between the two models after fine-tuning is shown in Figure 13. It can be observed that, due to concept balancing during the training process with IMBA loss, the IMBA distance after training with IMBA loss pays more attention to tail concepts (**red words**). Consequently, the IMBA distance in the corresponding regions (**green boxes**) is smaller compared to training with diffusion loss.

## C. More Experiment Results of the Model Size

When testing different model sizes on the same dataset in Section 3, we observed that even with significant differences in model size, the generated images exhibit highly similar structural features given the same initial noise and text prompts, as illustrated in Figure 14. This suggests that once a model reaches a certain size, the dataset itself becomes more influential in determining the generated images rather than the model capacity. Larger models indeed have


Figure 14. Generation results of models with different sizes from the same initial noise.

better convergence capabilities, but they do not dictate the high-dimensional semantics or concept composition abilities of the images.