ANALYZING THE ALIGNMENT–DIVERSITY TRADE-OFF IN TEXT-TO-IMAGE DIFFUSION MODELS

1 BACKGROUND: WHAT IS THE ALIGNMENT–DIVERSITY TRADE-OFF?

Text-to-image diffusion models have achieved impressive progress in generating realistic and semantically coherent images from textual descriptions. However, recent research shows that as models are optimized to produce outputs more faithful to textual prompts, their visual diversity tends to decline — a phenomenon known as the **alignment–diversity trade-off**. This trade-off reflects a fundamental tension between **textual accuracy** and **sample variability**, as strong alignment often constrains the sampling space and leads to visually repetitive results Jena et al. (2025); Miao et al. (2024); Li et al. (2024).

The problem is not limited to specific training strategies but emerges across alignment paradigms such as reward-based optimization, preference tuning, and reinforcement-guided fine-tuning. While these methods enhance prompt faithfulness, they frequently lead to *mode collapse* or reduced variability among generated samples. Recent works Jena et al. (2025); Miao et al. (2024) reveal that the balance between reward intensity, entropy regularization, and diversity control plays a key role in this trade-off. Moreover, broader surveys such as Li et al. (2024) argue that developing alignment mechanisms that do not sacrifice diversity is a critical step toward human-aligned and creatively capable diffusion models.

Prerequisite Knowledge Basic familiarity with: (i) diffusion-based generative models, (ii) reinforcement or reward-based optimization, (iii) evaluation metrics for alignment and diversity such as CLIPScore, LPIPS, SSIM, and Shannon entropy.

2 OUR APPROACH / DIRECTION

This project aims to systematically analyze how different alignment paradigms influence **output diversity** in text-to-image diffusion models. Instead of proposing a new alignment method, the goal is to understand and characterize the mechanisms that govern the loss of diversity as alignment pressure increases.

The analysis will focus on:

- Quantitatively evaluating how increasing alignment strength (e.g., through stronger rewards or loss weighting) affects diversity metrics such as LPIPS and image entropy;
- Exploring whether regularization techniques (e.g., entropy-based or KL penalties) can mitigate the diversity loss;
- Developing a unified evaluation framework to measure both alignment and diversity consistently across different model settings.

This investigation aims to reveal whether an optimal trade-off point exists where models can achieve high alignment without suffering a collapse in diversity.

3 RESEARCH QUESTIONS

The following questions guide our study:

- How does alignment optimization quantitatively affect diversity across diffusion models?

- Can regularization or entropy-aware objectives preserve or enhance diversity without weakening alignment?
- What measurable relationship exists between alignment metrics and diversity indicators?
- Is there a regime where both alignment and diversity can be jointly maximized?
- How can these findings inform future alignment frameworks for text-to-image generation?
- Can a more precise and reliable metric be proposed to measure visual diversity, and do existing metrics (LPIPS, SSIM, entropy) accurately capture the perceptual variability of generated images?

REFERENCES

Pratyush Jena, Neeraj Singh, Mayank Agarwal, and Rishabh Arora. Elucidating optimal reward–diversity trade-offs in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Jena_Elucidating_Optimal_Reward-Diversity_Tradeoffs_in_Text-to-Image_Diffusion_Models_WACV_2025_paper.html.

Ziang Li, Zhenguo Li, Chen Wang, and Xihui Liu. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024. URL https://arxiv.org/abs/2409.07253.

Jiaxu Miao, Hang Xu, Xiaotian Li, Wei Xu, and Dacheng Tao. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Miao_Training_Diffusion_Models_Towards_Diverse_Image_Generation_with_Reinforcement_Learning_CVPR_2024_paper.html.