# Supplementary Material

## A. Empirical Studies on Finetuning Text Encoder in Stable Diffusion Model

To isolate the impact of text-encoder fine-tuning, we fine-tune the CLIP text encoder (together with the UNet) using only the standard reconstruction loss. Table 1 reports YOLOv11 precision, recall, F1-score, and FID on the VFN dataset for several related works, with and without CLIP fine-tuning. Integrating text-encoder fine-tuning yields substantial gains. In contrast, freezing the text encoder while fine-tuning only the UNet produces limited performance. These findings confirm that the pretrained CLIP encoder lacks sufficient food-domain knowledge, and even vanilla fine-tuning is essential for generating more accurate, recognizable food images. Note that Stable Diffusion is pretrained on the LAION-5B dataset [7] with limited representation of food images, which usually have many multi-noun categories.

## B. Quantitative Results across all categories in two datasets

Table 2 reports YOLOv11 precision, recall, F1-score, and FID on the VFN and UEC-256 datasets, with real-image detection rates serving as an Reference Value. Our method, FoCULR, consistently surpasses all baselines, which means multi-noun category image generation is a critical issue in a food dataset and also our proposed method does not affect the performance from other categories.

Table 3 reports precision, recall, F1-score, and FID on the full VFN dataset for three configurations: CFIG only (no text-encoder fine-tuning), FDALA only (fine-tuning without negative-prompt scheduling), and the combined FDALA+CFIG method. Applying CFIG alone raises precision modestly by suppressing redundant elements, but recall remains low and FID high. FDALA alone yields substantial gains compared with CFIG only, demonstrating that text-encoder fine-tuning is critical for semantic alignment. The full FDALA+CFIG combination further boosts all metrics, confirming that negative-prompt scheduling refines layouts beyond fine-tuning alone.

## C. Ablation studies on the number of timesteps to apply negative prompts in CFIG (VFN dataset)

We perform an ablation study to investigate the impact of varying the number of timesteps during which negative prompts are applied within the CFIG module (Table 5). Negative prompts in the early denoising steps help to suppress redundant layout initialization, thus reducing the likelihood of unwanted food objects appearing in the generated image.

Our results indicate that applying negative prompts for too many timesteps can degrade generation quality, as prolonged suppression inhibits the generation of fine-grained details during the later stages of denoising. Conversely, using negative prompts for too few timesteps is insufficient to effectively exclude redundant object layouts. Among the tested configurations, applying negative prompts for 5 timesteps ($t_{\text{threshold}} = 5$) achieves the best balance between avoiding redundant layout initialization and preserving fine-grained detail, thus resulting in optimal performance.

## D. Effect of Prompt Rewrite

We also conduct empirical studies on the effect of prompt rewrite, as shown in Figure 1. With more information provided in the prompt—for example, explicitly stating that the food object appears in a single dish—the generated images still contain redundant elements (e.g., extra corn beside a corn dog, or corns replacing the intended "cheese corn snack"). This indicates that simply rephrasing the prompt does not fundamentally resolve the multi-noun challenge. Although Stable Diffusion 3 incorporates both a CLIP encoder and a T5 encoder, neither is explicitly trained to model syntactic head–modifier structures within food categories. As a result, the encoders still tend to overemphasize individual tokens (e.g., "corn") as separate visual entities rather than attributes linked to the head noun (e.g., "dog" or "snack"). Hence, while prompt rewriting enriches the textual input, it cannot prevent redundant or misplaced elements in the generated images.

Table 1. YOLOv11 detection and FID score results on generated images for VFN dataset across all categories for emprical studies on finetuning text encoder in stable diffusion model

| Generation Method | No fine-tuning on CLIP text encoder | | | | CLIP text encoder fine-tuned | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision↑ | Recall↑ | F-1 score↑ | FID score↓ | Precision↑ | Recall↑ | F-1 score↑ | FID score↓ |
| Real images(Reference Value) | 0.743 | 0.777 | 0.76 | – | 0.743 | 0.777 | 0.76 | – |
| Stable diffusion [4] | 0.646 | 0.677 | 0.661 | 38.6 | 0.782 | 0.832 | 0.806 | 32.8 |
| Structured diffusion [2] | 0.635 | 0.693 | 0.663 | 37.4 | 0.661 | 0.707 | 0.683 | 37.0 |
| Syngen [6] | 0.499 | 0.465 | 0.481 | 42.1 | 0.678 | 0.7 | 0.689 | 32.1 |

Table 2. Comparison with related works on generated images for VFN and UEC-256 dataset across all categories.

| Method | Text encoder fine-tuning ? | VFN dataset | | | | UEC-256 dataset | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Precision ↑ | Recall↑ | F-1 score↑ | FID score↓ | Precision↑ | Recall↑ | F-1 score↑ | FID score↓ |
| Real images (Reference Value) | – | 0.743 | 0.777 | 0.76 | – | 0.762 | 0.725 | 0.743 | – |
| Stable diffusion [4] | ✗ | 0.646 | 0.677 | 0.661 | 38.6 | 0.299 | 0.116 | 0.167 | 27.0 |
| Structured diffusion [2] | ✗ | 0.635 | 0.693 | 0.663 | 37.4 | 0.247 | 0.097 | 0.139 | 28.7 |
| Syngen [6] | ✗ | 0.499 | 0.465 | 0.481 | 42.1 | 0.121 | 0.05 | 0.071 | 52.6 |
| Stable diffusion 3 (SD3) [1] | ✗ | 0.681 | 0.689 | 0.685 | 42.8 | 0.306 | 0.309 | 0.307 | 34.9 |
| SD3 + Prompt Rewrite [3] | ✗ | 0.663 | 0.689 | 0.676 | 43.5 | 0.307 | 0.296 | 0.301 | 35.4 |
| TextCraftor [5] | ✓ | 0.798 | 0.848 | 0.822 | 34.4 | 0.653 | 0.263 | 0.375 | **25.4** |
| FoCULR(Ours) | ✓ | **0.813** | **0.872** | **0.841** | **33.0** | **0.71** | **0.279** | **0.4** | 25.9 |

Table 3. Ablation studies of our method for VFN dataset across all categories

| FDALA | CFIG | Text encoder fine-tuning? | Precision↑ | Recall↑ | F-1 score↑ |
|---|---|---|---|---|---|
| ✗ | ✓ | ✗ | 0.667 | 0.704 | 0.685 |
| ✓ | ✗ | ✓ | 0.811 | 0.86 | 0.835 |
| ✓ | ✓ | ✓ | **0.813** | **0.872** | **0.841** |

Table 4. Ablation studies of our method for UEC-256 dataset across all categories

| FDALA | CFIG | Text encoder fine-tuning? | Precision↑ | Recall↑ | F-1 score↑ |
|---|---|---|---|---|---|
| ✗ | ✓ | ✗ | 0.327 | 0.118 | 0.173 |
| ✓ | ✗ | ✓ | 0.689 | 0.278 | 0.396 |
| ✓ | ✓ | ✓ | **0.71** | **0.279** | **0.4** |

Table 5. Ablation studies on timesteps for applying negative prompts in CFIG on VFN dataset

| Total inference steps | Negative Prompt steps | Precision | Recall | F-1 score |
|---|---|---|---|---|
| **Real images (Reference Value)** | | 0.347 | 0.312 | 0.329 |
| **100 No negative prompt** | 0 | 0.451 | 0.432 | 0.441 |
| **50** | 3 | 0.432 | 0.416 | 0.424 |
| **50** | 5 | **0.457** | **0.433** | **0.445** |
| **50** | 10 | 0.417 | 0.392 | 0.404 |
| **50** | 15 | 0.437 | 0.397 | 0.416 |
| **50** | 20 | 0.4 | 0.375 | 0.387 |

## References


[1] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *Forty-first international conference on machine learning*, 2024.

[2] W. Feng, X. He, T. Fu, V. Jampani, A. Reddy Akula, P. Narayana, S. Basu, X. Wang, and W. Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *The Eleventh International Conference on Learning Representations*, 2023.

[3] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross-attention control. *Internal Conference on Learning Representation*, 2023.

[4] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

[5] Y. Li, X. Liu, A. Kag, J. Hu, Y. Idelbayev, D. Sagar, Y. Wang,

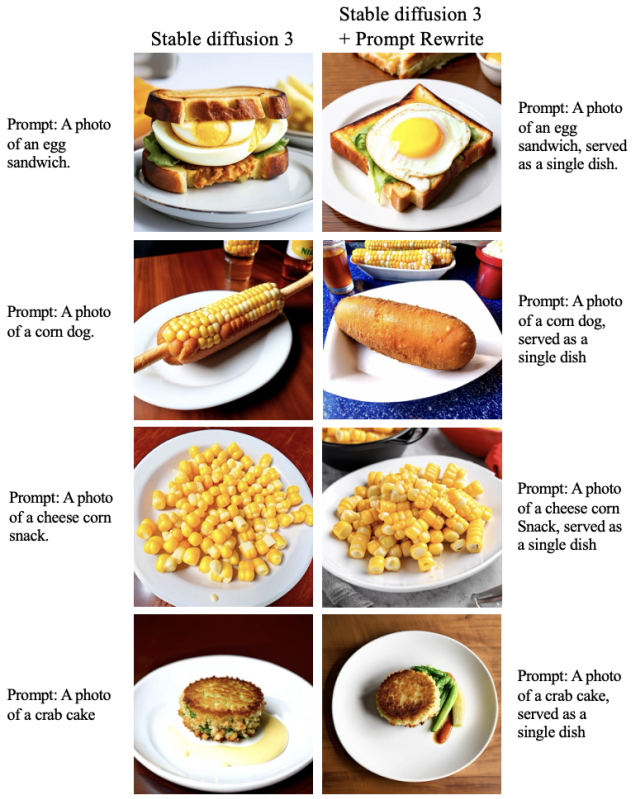


Figure 1. Empirical studies on the effect of prompt rewrite

S. Tulyakov, and J. Ren. Textcraftor: Your text encoder can be image quality controller. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7985–7995, 2024.

[6] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 2024.

[7] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, pages 25278–25294, 2022.