## A. Text-to-Image Qualitative Results

We visualize generations between our REPA-MMDiT models described in Section 5.3 trained with flow matching (FM) loss and with ΔFM on CC3M with a batch size of 256 for 400K iterations in Figure 6. We plot images in pairs, with FM images on the left and ΔFM images on the right, and show the respective caption for each pair above. All images are generated without classifier-free guidance and using NFE=50, and are the same images used in Table 3.

## B. Deriving Contrastive-Flow Matching Interference

### B.1. Closed-form solution to Eq. 4

We first re-introduce Eq. 4 for convenience,

$$\mathcal{L}^{(\Delta\text{FM})}(\theta) = \mathrm{E}\begin{bmatrix} ||v_\theta(x_t,t,y) - (\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon)||^2 \\ -\lambda||v_\theta(x_t,t,y) - (\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})||^2 \end{bmatrix}$$

Minimizing the expectation, expanding all norms and letting $v(\theta) = v(x_t,t,y)$, we can simplify the expectation to:

$$= \min_\theta \mathrm{E}\begin{bmatrix} (1-\lambda)v(\theta)^T v(\theta) \\ -2v(\theta)^T\left[(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon) - \lambda(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})\right] \\ +(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon)^T(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon) \\ -\lambda(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})^T(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon}) \end{bmatrix} \tag{7}$$

$$= \min_\theta \mathrm{E}\begin{bmatrix} (1-\lambda)v(\theta)^T v(\theta) \\ -2v(\theta)^T\left[(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon) - \lambda(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})\right] \end{bmatrix} \tag{8}$$

$$\overset{\propto}{\sim} \min_\theta \mathrm{E}\left[\left\|\begin{matrix} \sqrt{1-\lambda}v(\theta) \\ -\dfrac{(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon) - \lambda(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})}{\sqrt{1-\lambda}} \end{matrix}\right\|_2^2\right] \tag{9}$$

Setting the gradient with respect to $v(\theta)$ to 0,

$$\sqrt{1-\lambda}v(\theta)^* = \mathrm{E}\left[\frac{(\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon) - \lambda(\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon})}{\sqrt{1-\lambda}}\right] \tag{10}$$

$$v(\theta)^* = \frac{\mathrm{E}\left[\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon\right] - \lambda\mathrm{E}\left[\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon}\right]}{1-\lambda} \tag{11}$$

Finally, observe that $\mathrm{E}\left[\dot{\alpha}_t\hat{x} + \dot{\sigma}_t\epsilon\right]$ is the solution to the flow matching objective. Setting $\mathrm{E}\left[\dot{\alpha}_t\tilde{x} + \dot{\sigma}_t\tilde{\epsilon}\right] = \hat{T}$ and observing that $x_t$ does not depend on $\hat{x}$ or $\hat{\epsilon}$ we obtain:

$$\min_\theta \mathcal{L}^{(\Delta\text{FM})}(\theta) = \frac{\min_\theta \mathcal{L}^{(\text{FM})}(\theta) - \lambda\hat{T}}{1-\lambda} \tag{12}$$

### B.2. Coupling with CFG

Classifier-free guidance (CFG) is originally defined over the flow matching solution of $\min_\theta \mathcal{L}^{(\text{FM})}$. Re-writing Eq. 12

| Model | Batch Size | Metrics FID ↓ | IS ↑ | sFID ↓ |
|---|---|---|---|---|
| REPA SiT-B/2 | 256 | 27.33 | 61.60 | 11.70 |
| + Using ΔFM | 256 | **20.52** | **69.71** | **5.47** |
| REPA SiT-B/2 | 512 | 24.45 | 69.15 | 11.42 |
| + Using ΔFM | 512 | **17.06** | **81.41** | **5.29** |
| REPA SiT-B/2 | 1024 | 22.00 | 76.15 | 11.76 |
| + Using ΔFM | 1024 | **15.23** | **88.53** | **5.20** |
| REPA SiT-XL/2 | 256 | 11.14 | 115.83 | 8.25 |
| + Using ΔFM | 256 | **7.29** | **129.89** | **4.93** |
| REPA SiT-XL/2 | 512 | 10.15 | 129.43 | 9.00 |
| + Using ΔFM | 512 | **6.36** | **146.17** | **5.42** |

Table 6. **ΔFM Scales with Batch Size.** We train all models for 400K iterations and strictly follow the protocol of [44]. All metrics are measured with the SDE Euler-Maruyama sampler with NFE=50 and without classifier guidance. We use $\lambda = 0.05$ for all models trained with ΔFM and do not change any other hyperparameters. ↑ indicates that higher values are better, with ↓ denoting the opposite. Improvement using ΔFM evenly scales with batch-size, and even outperforms flow-matching models with *half* the batch-size.

and substituting it into the CFG equation, we obtain:

$$CFG = wv^{(FM)}(x_t,t,y) + (1-w)v^{(FM)}(x_t,t,\emptyset) \tag{13}$$

$$= \begin{bmatrix} w\left[(1-\lambda)v^{(\Delta\text{FM})}(x_t,t,y) + \lambda\hat{T}\right] \\ -(1-w)\left[(1-\lambda)v^{(\Delta\text{FM})}(x_t,t,\emptyset) + \lambda\hat{T}\right] \end{bmatrix} \tag{14}$$

$$= \left[(1-\lambda)\begin{bmatrix} wv^{(\Delta\text{FM})}(x_t,t,y) \\ +(1-w)v^{(\Delta\text{FM})}(x_t,t,\emptyset) \end{bmatrix} + \lambda\hat{T}\right] \tag{15}$$

Letting $v(x_t|y) = v^{(\Delta\text{FM})}(x_t,t,y)$ and $v(x_t|\emptyset) = v^{(\Delta\text{FM})}(x_t,t,\emptyset)$, we obtain the Eq. from Section 5.4: $\hat{CFG} = (1-\lambda)\left[wv(x_t|y) + (1-w)v(x_t|\emptyset)\right] + \lambda\hat{T}$.

### B.3. Other CFG Couplings

While we find that our proposed coupling strategy for ΔFM and CFG works well for our setting, other suitable variations may also exist. For instance, one may instead reduce conflicts by following the equation: $C\tilde{F}G = (w+\lambda)v(x_t|y) - (1-w)v(x_t|\emptyset) - \lambda\hat{T}$, where $\lambda$, and $w$ are free hyperparameters. We leave such exploration to future work.

## C. Effects of batch size on ΔFM.

In Table 6, we study the effects of batch size on our loss. It is well known that batch size has an important effect on contrastive style losses [5, 7, 15] that draw negatives within the batch. This can be understood as a sample diversity issue. If the batch size is larger than negative samples within the batch are more representative of the true distribution. In this table, we see a similar trend: larger batch sizes are important

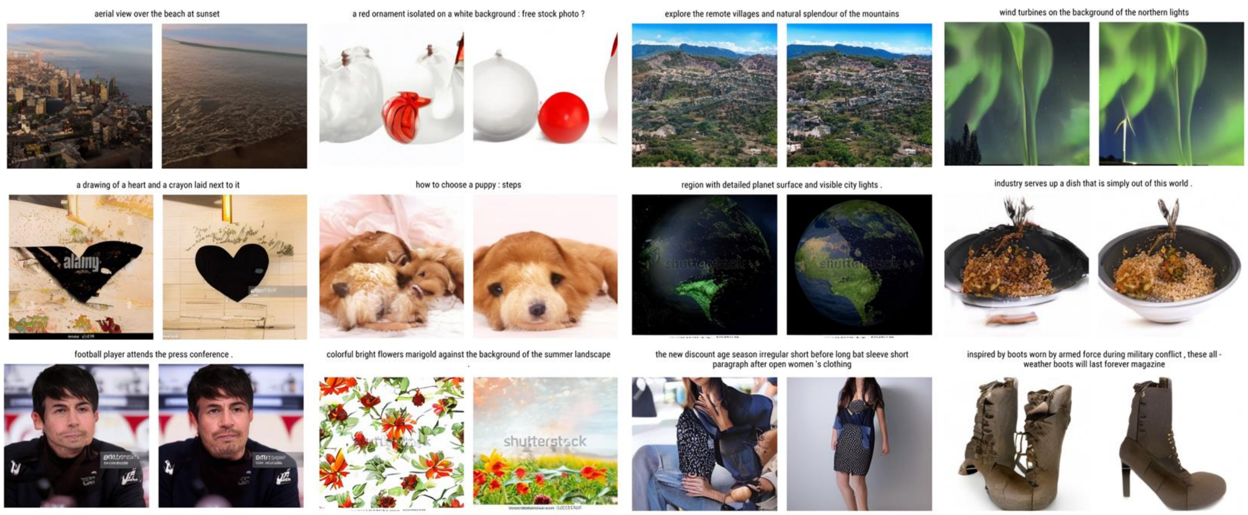


Figure 6. **CC3M side-by-side generations between a REPA-MMDiT model trained with flow matching (left) and △FM (right).** Models are trained for 400K iterations using a batch-size of 256 and images are generated without classifier-free guidance and using NFE=50.

for maximizing the performance of ΔFM across several model scales. We also maintain our improvements over the REPA baseline through all batch sizes and model scales.