Guiding Token-Sparse Diffusion Models

Supplementary Material

A. Implementation Details

A.1. Training Details for T2I

Architecture We implement our transformer models [15, 71] largely following the Llama architecture [70]. In particular, we apply pre-normalization via RMSNorm [75], exclude bias parameters from all linear transformations, and employ rotary positional embeddings [69] in an axial configuration following the approach of Crowson et al. [12]. The feedforward network (FFN) design mirrors that of Llama, utilizing the SwiGLU activation [63] and an expansion ratio of $\frac{8}{3}$.

Model We train a modern T2I diffusion transformer with 2.5B parameters. To apply TREAD [33], we mask tokens and positional indices simultaneously and reintroduce them at layer 30. We use Internvl3-2B [78] as the text encoder. In addition, we incorporate insights from Ma et al. [43], specifically employing two TransformerLayers after the frozen VLM and using a general system prompt as a prefix to our captions: `"Describe the image by detailing the color, shape, size, texture, quantity, text, and spatial relationships of the objects."`. For more details on the model refer to Table A1.

Data We use InternVL3-2B [78] to recaption a 100M-sample subset of COYO-700M [7], producing four captions per image. First, we generate a highly detailed description of the image and then progressively distill it into three additional levels: multi-sentence descriptions, single-sentence descriptions, and finally keyword-level summaries. For the last three, we use the language capacity of the VLM exclusively to cut down on cost. After a first training stage, we filter the COYO subset by aesthetics score (>5) and add synthetic data from JourneyDB [51] and Flux-6M [17].

A.2. Hyperparameters for ImageNet

Unless stated otherwise we inherit the DiT [52] setting: AdamW [42], a fixed learning rate of 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, `bf16` precision, and latent-space training with the `stabilityai/sd-vae-ft-ema` VAE [56]. When we finetune LR is dropped to 10^{-5}. For routing and masking specific parameters refer to Table A2.

B. Experiment Details

B.1. Sparse Guidance in ImageNet

$\mathbf{SG}_{\mathbf{FLOPS}}$ from Section 4.2 is obtained using the same checkpoint for the high capacity and low capacity model.

Hyperparameter	**TR-DiT-2.5B**
Optimizer	
Batch size	3,072
Optimizer	AdamW
Learning rate	5×10^{-5}
(β_1, β_2)	(0.9, 0.95)
Architecture	
Embedding dim	2,048
Attention heads	16
Transformer layers	34
TREAD settings	
Route	$\mathbf{r}_{2\to30}$
Selection ratio	0.5

Table A1. Hyperparameter setup for our TR-DiT-2.5B model and the TREAD routing schedule.

Hyperparameter	**Routing**	**Masking**
Optimizer		
Batch size	256	256
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}
(β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)
Finetune		
Batch size	256	256
Learning rate	1×10^{-5}	1×10^{-5}
Architecture		
Embedding dim	1,152	1,152
Attention heads	16	16
Transformer layers	28	28
TREAD settings		
Route	$\mathbf{r}_{2\to24}$	–
Selection ratio	0.5	–
MaskDiT settings		
D^{dec}Embedding dim	–	512
D^{dec}Attention heads	–	16
D^{dec}Transformer layers	–	8
Selection ratio	–	0.5

Table A2. Hyperparameter setup for the XL/2 backbones with additional information for routing [33] and masking [76] methods. D^{dec} refers to the decoder head placed upon the normal DiT-XL/2. $\mathbf{r}_{2\to24}$ refers to the route from layer 2 to layer 24.

Both are conditional and the distribution discrepancy is created solely via different routing rates. We find $\gamma_{\text{strong}} = 0.5, \gamma_{\text{weak}} = 0.9$ to achieve good FID while substantially decreasing FLOPS.

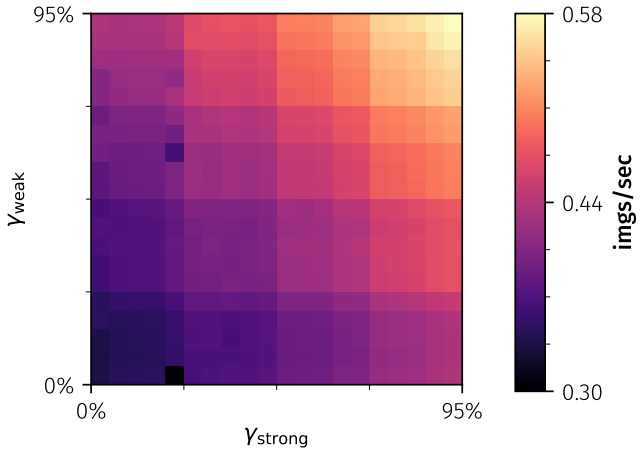

Figure A1. Inference speed for the guided setting. Lower left corner with zero γ_{strong}, γ_{weak} resembles naive guided inference. Introducing sparsity (Sparse Guidance) allows for drastically improved throughput showcased by brighter colors towards the top right corner.

SG_{FID} (see Section 4.2, Table 3) is obtained through the usage of an early checkpoint of the same model training run. More specifically, we utilize a checkpoint with 50k training iterations. Furthermore, we apply cosine decay from 0.6 to 0.0 on the auxiliary model and the inverse on the main model. This aligns with the findings from Figure 7 where $\gamma_{\text{strong}}, \gamma_{\text{weak}}$ can be used to make up for undertrained auxiliary models. We achieve similar FID with other checkpoints and adjusted routing rates.

B.2. Sparse Guidance in Large Scale T2I Models

In Table 4 we show that applying our proposed Sparse Guidance to scaled T2I models yields better performance than CFG. Additionally, Sparse Guidance enables faster inference as seen in Figure A1 where a grid over the γ_{strong},γ_{weak} with a 0.05 stepsize is shown.

GenEval [19] For GenEval (see Table 5), we stack our proposed Sparse Guidance method on top of Classifier-free Guidance and utilize $\omega = 2.5, \gamma_{\text{strong}} = 0.2$ and $\gamma_{\text{weak}} = 0.7$.

HPSv3 [48] For the HPSv3 score (see Table 4), we follow the proposed benchmark in Ma et al. [48] with identical prompts. We utilize Sparse Guidance with $\omega = 1.8, \gamma_{\text{strong}} = 0.1$ and $\gamma_{\text{weak}} = 0.8$.

C. Auxiliary MAE loss under Flow Matching

To facilitate a fair comparison between our SiT [44] baseline and MaskDiT [76], we derive the MaskedAutoEncoder (MAE) loss for the flow-matching objective (see Table 1, Figure 6). MaskDiT [76] combines a score-matching loss on visible tokens with a masked reconstruction (MAE) objective on masked tokens in diffusion models. We generalize this formulation to the *flow-matching* objective. Let $\mathcal{I}$ denote the token index set and $\mathbf{M} \in \{0,1\}^{\mathcal{I}}$ a random binary mask (1 for masked, 0 for visible). We define the visible mask as $\bar{\mathbf{M}} = \mathbf{1} - \mathbf{M}$. Following [76], the masked reconstruction loss is:

$$\mathcal{L}_{\text{MAE}} = \mathbb{E}_{x\sim p_{\text{data}}}\mathbb{E}_{t\sim[0,1]}\mathbb{E}_{\mathbf{M}}\big\|\big(D_\theta(x_t \odot \bar{\mathbf{M}}, t) - x\big) \odot \mathbf{M}\big\|^2, \tag{A1}$$

where D_θ predicts the denoised image at time t and $\odot$ denotes the Hadamard product. Unlike diffusion models, which predict the score $\nabla_{x_t} \log p_t(x_t)$, flow matching directly parameterizes the instantaneous displacement of particles along this trajectory. Given the path definition in Eq. 1, the latent states satisfy

$$x - x_t \;=\; (1-t)(x-z) \;=\; (1-t)\, v^\star(x_t, t), \tag{A2}$$

where $v^\star(x_t, t)$ is the oracle velocity field driving the transformation from z to x. This relation reveals that reconstructing a future state x_t from a clean sample x is equivalent to estimating the target velocity $v^\star(x_t, t)$ up to the scalar factor $(1-t)$. Hence, in the flow-matching formulation, masked reconstruction can be interpreted as learning to predict the intermediate flow direction that transports partially visible tokens toward their clean targets. Replacing $v^\star$ by its learned approximation v_θ, we have

$$D_\theta(x_t, t) - x_t \approx (1-t)\, v_\theta(x_t, t).$$

Consequently, the masked reconstruction term restricted to masked tokens can be reformulated as:

$$\begin{aligned}\mathcal{L}_{\text{MAE}} &= \mathbb{E}_x\mathbb{E}_{t\sim[0,1]}\mathbb{E}_{\mathbf{M}}\big\|(1-t)\, v_\theta(x_t \odot \bar{\mathbf{M}}, t) \odot \mathbf{M}\big\|^2 \\ &= \mathbb{E}_x\mathbb{E}_{t\sim[0,1]}\mathbb{E}_{\mathbf{M}}(1-t)^2\big\|v_\theta(x_t \odot \bar{\mathbf{M}}, t) \odot \mathbf{M}\big\|^2.\end{aligned} \tag{A3}$$

The overall training objective combines the standard flow-matching loss with the auxiliary masked reconstruction term. According to [33], routing models do not require additional auxiliary losses, so we use the standard flow matching objective. The final loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{FM-mask}} = \mathbb{E}_{x,z,t}\Big[\big\|\,\bar{\mathbf{M}} \odot \big(v_\theta(x_t, t) - v^\star(x_t, t)\big)\,\big\|_2^2 \\ + \lambda\, \mathbb{E}_{x,t,\mathbf{M}}(1-t)^2\big\|v_\theta(x_t \odot \bar{\mathbf{M}}, t) \odot \mathbf{M}\big\|_2^2\Big],\end{aligned} \tag{A4}$$

where λ balances the contribution of the masked reconstruction objective. In practice, we set λ empirically to ensure comparable magnitudes of the gradient between the two terms.

D. Guidance Interaction Exploration

Using the notation from Sec. 3.2, let

$$D_s(c) := D_\theta(x_t, t, c; \gamma_{\text{strong}}), \qquad D_w(c) := D_\theta(x_t, t, c; \gamma_{\text{weak}}),$$

with analogous definitions for $D_s(\emptyset)$ and $D_w(\emptyset)$. We compare four interaction strategies for combining CFG and SG.

A) Direct. We directly add the CFG direction and the SG direction:

$$\tilde{D}_\theta^{\text{direct}} = D_w(\emptyset) + w_{\text{cfg}}\big(D_s(c) - D_w(\emptyset)\big) + w_{\text{sg}}\big(D_s(c) - D_w(c)\big). \tag{A5}$$

B) Inner. Following the Inner-Guidance form, we use the jointly conditioned prediction and subtract the partially dropped branches:

$$\tilde{D}_\theta^{\text{inner}} = (1 + w_{\text{cfg}} + w_{\text{sg}})\, D_s(c) - w_{\text{cfg}}\, D_s(\emptyset) - w_{\text{sg}}\, D_w(c). \tag{A6}$$

C) Compositional. Following compositional guidance, we combine the condition-specific predictions additively around a shared unconditional branch:

$$\tilde{D}_\theta^{\text{comp}} = D_w(\emptyset) + w_{\text{cfg}}\big(D_s(c) - D_w(\emptyset)\big) + w_{\text{sg}}\big(D_w(c) - D_w(\emptyset)\big). \tag{A7}$$

D) IP2P-style. Following InstructPix2Pix, we apply the two signals sequentially, first SG and then CFG:

$$\tilde{D}_\theta^{\text{IP2P}} = D_w(\emptyset) + w_{\text{sg}}\big(D_w(c) - D_w(\emptyset)\big) + w_{\text{cfg}}\big(D_s(c) - D_w(c)\big). \tag{A8}$$

Among these interaction strategies, the direct formulation performs best, achieving the lowest FID while also requiring the fewest function evaluations per step.

Interaction Method	FID↓	NFE/step↓
A) direct	**2.14**	**2**
B) inner	10.70	3
C) compositional	11.14	3
D) IP2P-style	2.62	3

Table A3. Comparison of guidance interaction strategies on ImageNet.

E. Qualitative Samples

We provide additional qualitative text-to-image results in Figure A2 and Figure A3, where we directly compare Classifier-Free Guidance (CFG) with Sparse Guidance (SG) in our TREAD-DiT-2.5B. Complementing these comparisons, Figure A6 presents a broader selection of SG-generated outputs. All text-to-image samples are produced using prompts sourced from the HPSv3 [48] benchmark subset.

Subsequently, Figure A7 and Figure A8 display ImageNet-256 results, contrasting unguided predictions, AutoGuidance (AG), CFG, and our SG method. Finally, Figure A9, Figure A10, and Figure A11 offer uncurated qualitative comparisons between SG_{FID} and SG_{FLOPS} to illustrate their respective visual characteristics.

Figure A2. Qualitative examples comparing CFG to our proposed SG. Images with CFG tend to have more artifacts or seem blurry. SG provides crisp images with lower cost.

Figure A3. Qualitative examples comparing CFG to our proposed SG. Images with CFG tend to have more artifacts or seem blurry. SG provides crisp images with lower cost.

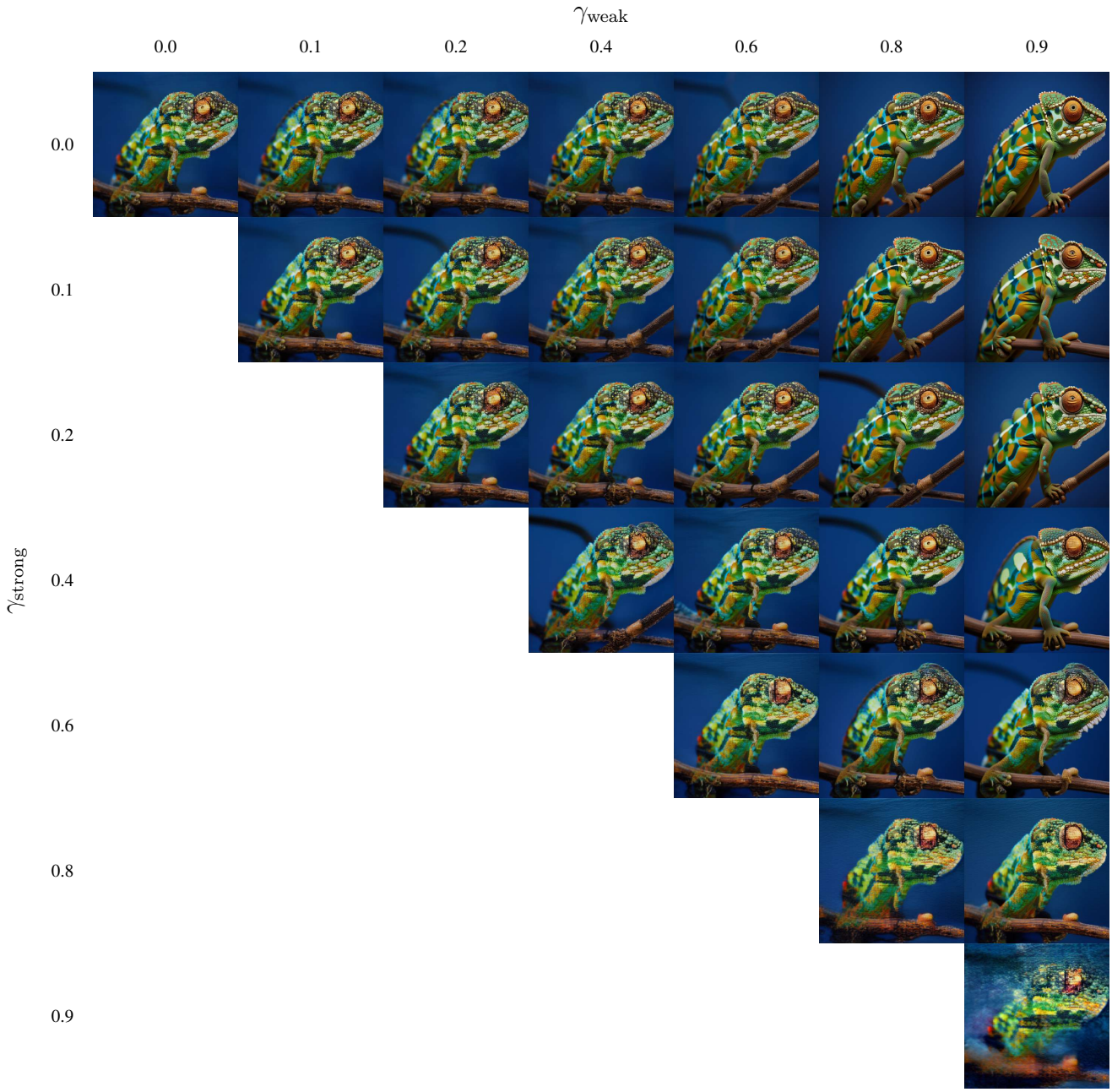

Figure A4. **Effect of Sparse Guidance on image quality.** Lower values for γ_{strong} and higher ones for γ_{weak} lead to best quality. *Prompt: The image showcases a vibrant chameleon perched on a slender branch against a striking, deep blue background. The chameleon is the main focus, presented in a close-up, side-view shot that emphasizes its fascinating textures and colors. Its skin is a tapestry of greens, blues, and browns, creating a mottled pattern of scales. Patches of white punctuate the green, and larger blotches of a reddish-brown add depth to its coloration. The head features a gradient of yellow and orange, drawing attention to its unique eye. The eye itself is a complex mix of orange and black. The texture of the chameleon's skin is visibly bumpy and scaled. Its curled toes are gripping the branch. The overall impression is one of natural beauty and intricate detail. The sharp focus and simple background allow for a detailed examination of the chameleon's unique features.*

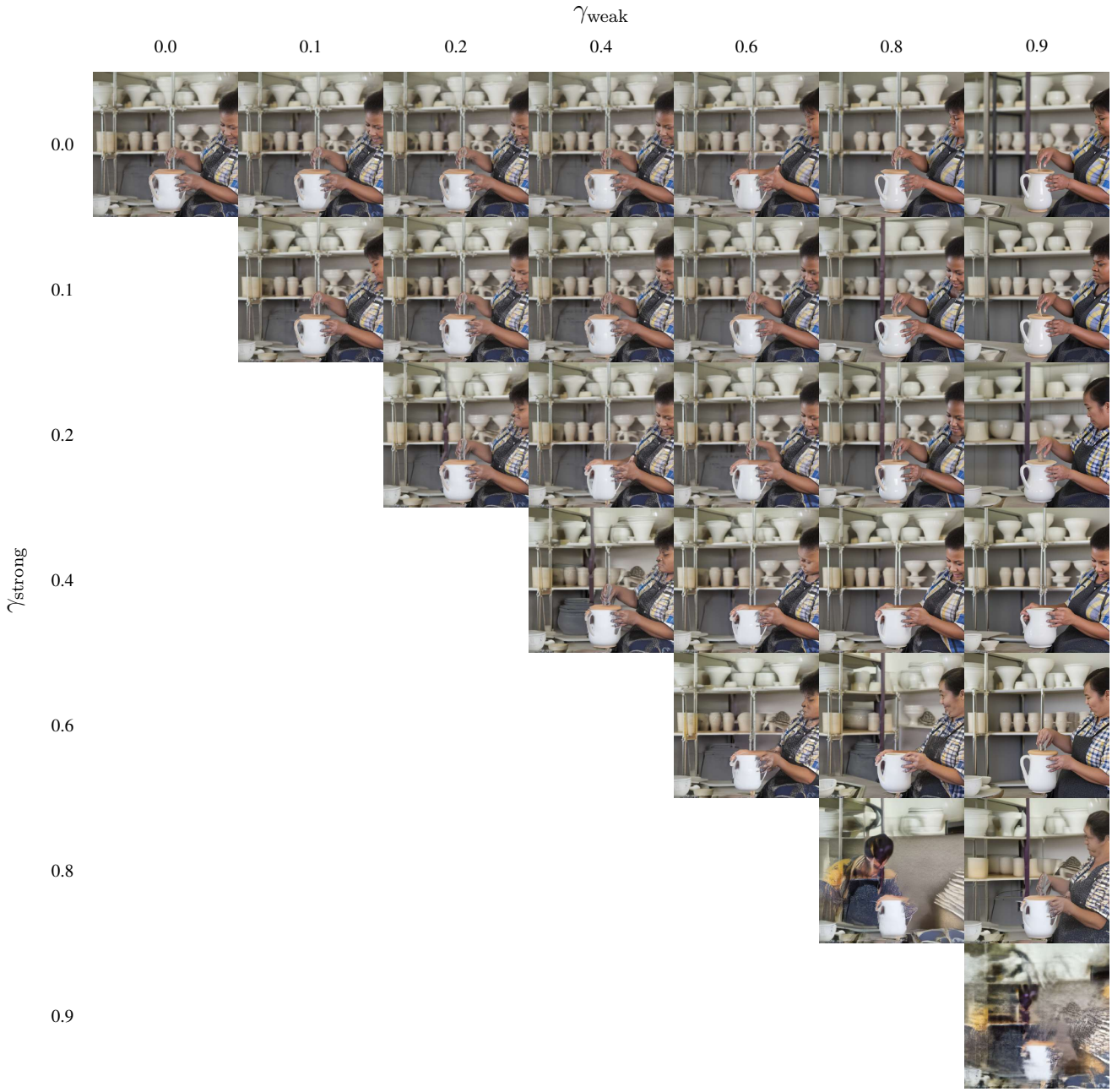

Figure A5. **Effect of Sparse Guidance on image quality.** Lower values for γ_{strong} and higher ones for γ_{weak} lead to best quality. *Prompt: The image shows a person engaged in the process of pottery making, specifically sanding a white ceramic pitcher. The individual, wearing a blue and yellow plaid shirt and a speckled black apron, carefully holds the pitcher with one hand while using a piece of sandpaper to smooth its surface with the other. In the background, a metal shelving unit filled with stacks of white ceramic bowls and cups indicates a workshop or studio setting. More unglazed ceramic pieces, including cups and plates, sit on a surface in the foreground, suggesting a production line or batch of pottery in progress. The lighting is soft and natural, highlighting the details of the ceramic surfaces and the craftsman's hands. The overall composition focuses on the tactile and meticulous nature of pottery making.*

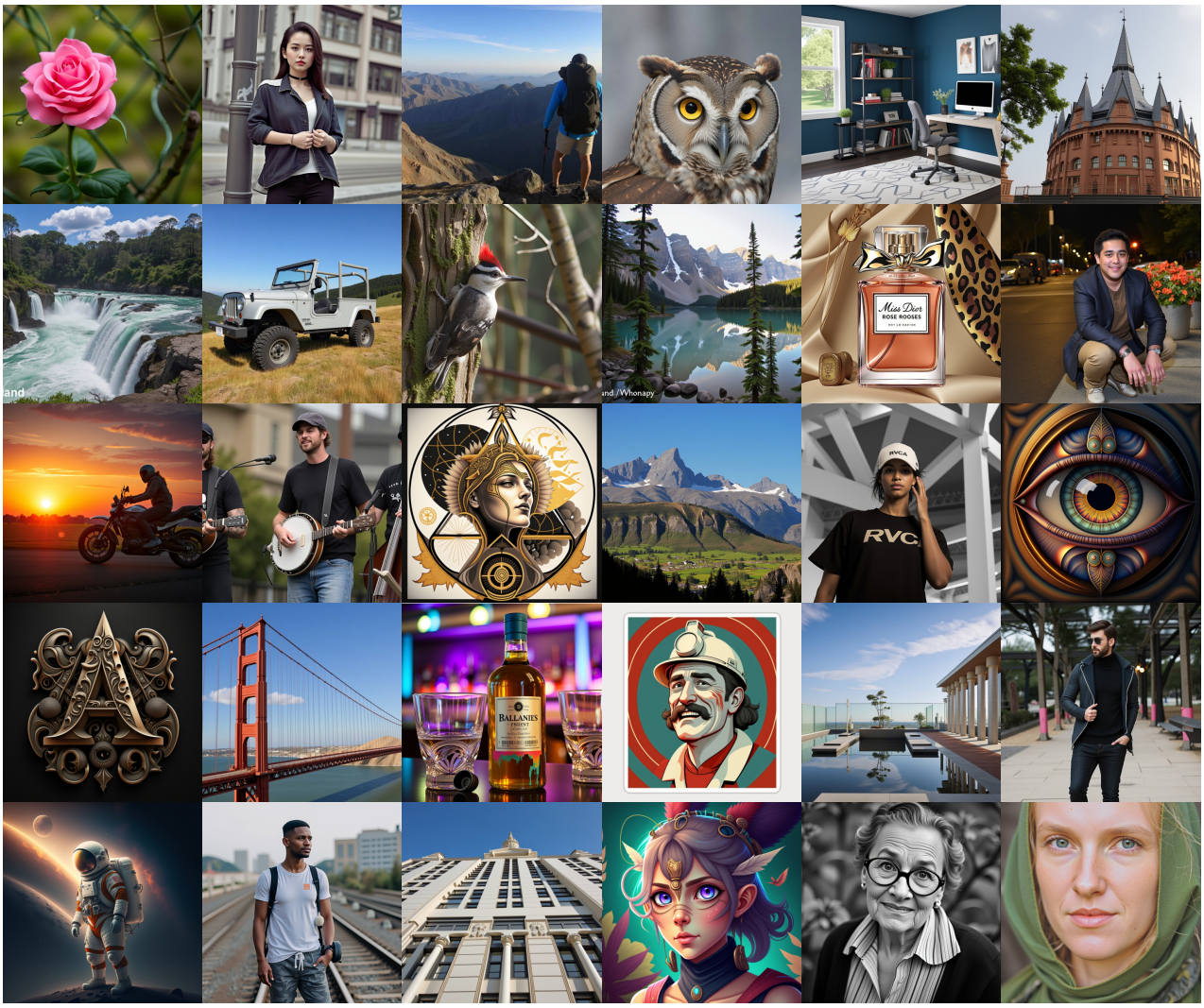

Figure A6. Additional samples generated using Sparse Guidance. Prompts are taken from the HPSv3 benchmark subset.

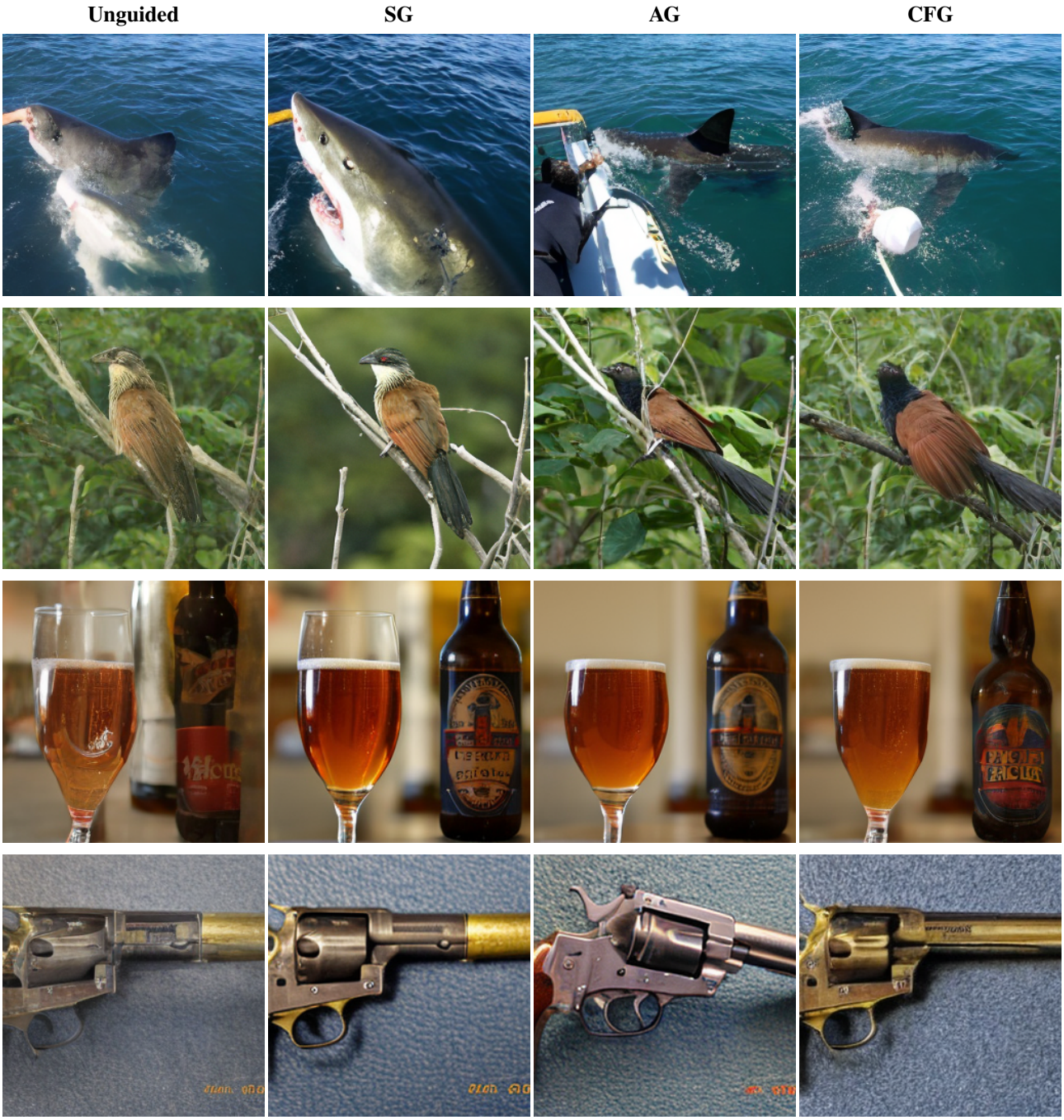

Figure A7. Qualitative samples using a guidance scale of $\omega = 2.5$ across different methods: Unguided, Sparse Guidance (SG), AutoGuidance (AG), and Classifier-Free Guidance (CFG).

Figure A8. Qualitative samples using a guidance scale of $\omega = 2.5$ across different methods: Unguided, Sparse Guidance (SG), AutoGuidance (AG), and Classifier-Free Guidance (CFG).

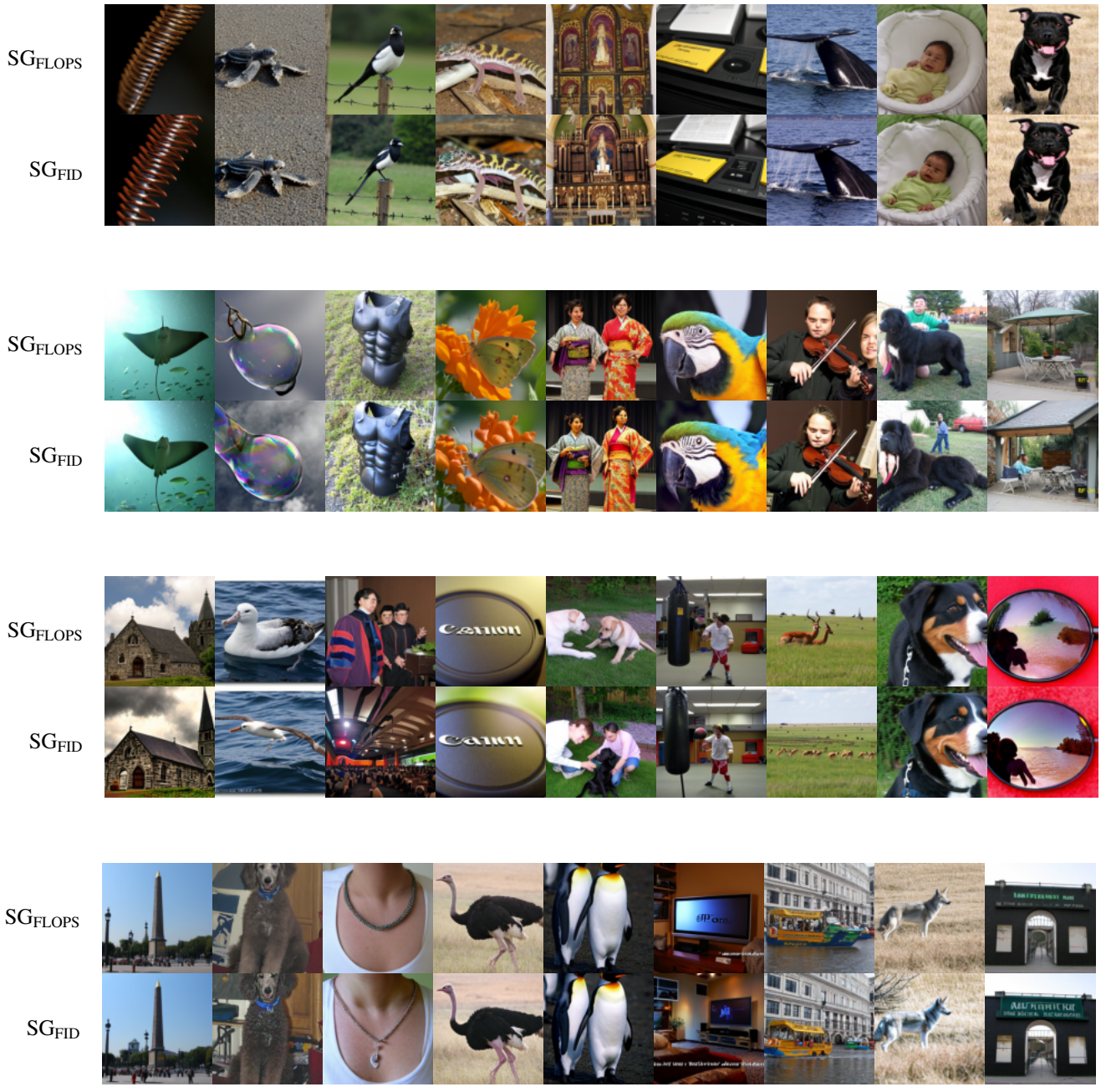

Figure A9. **Uncurated samples** of SG_{FLOPS} (top) and SG_{FID} (bottom) using $\omega = 2.5$.

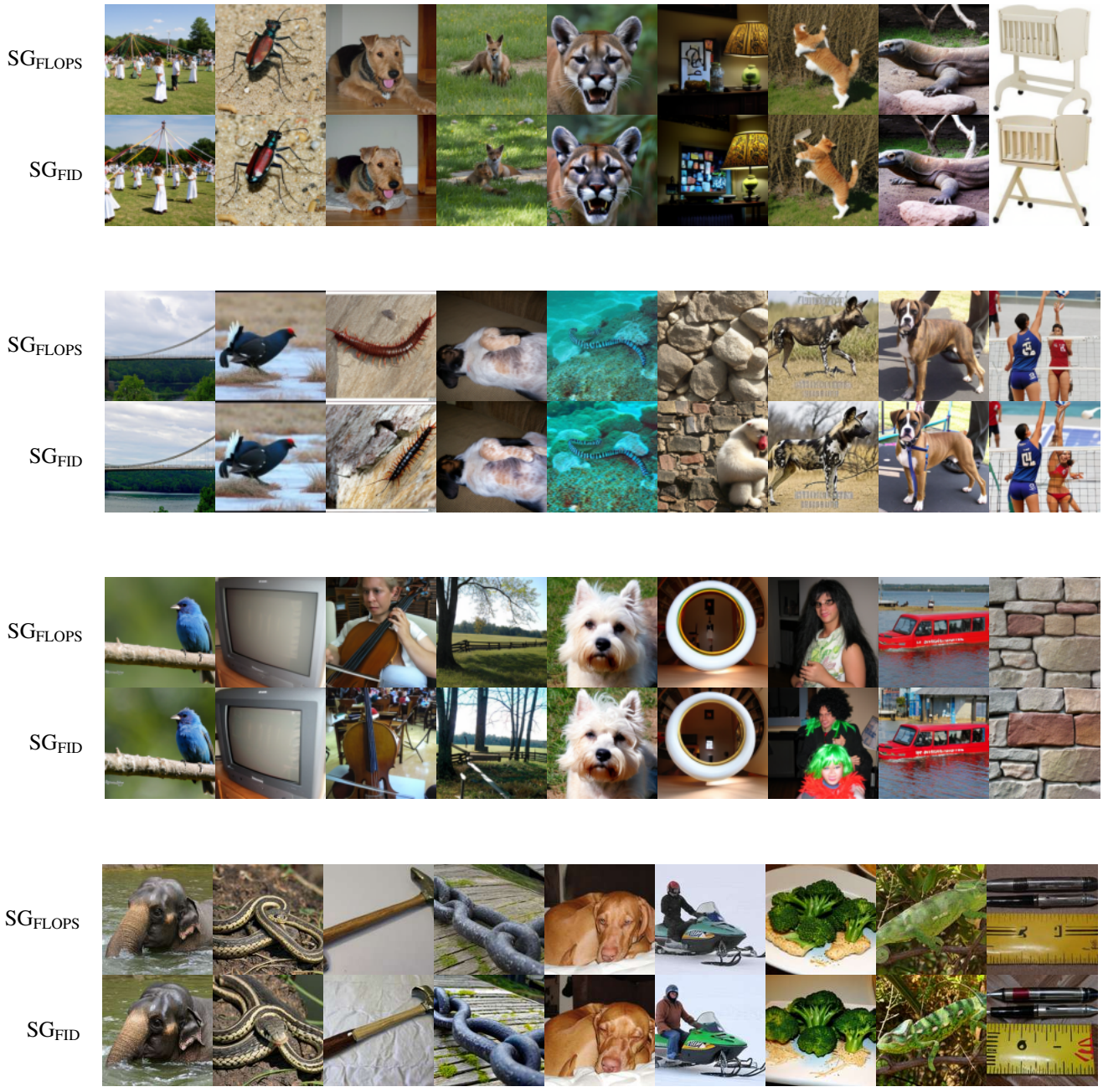

Figure A10. **Uncurated samples** of SG_{FLOPS} (top) and SG_{FID} (bottom) using $\omega = 2.5$.

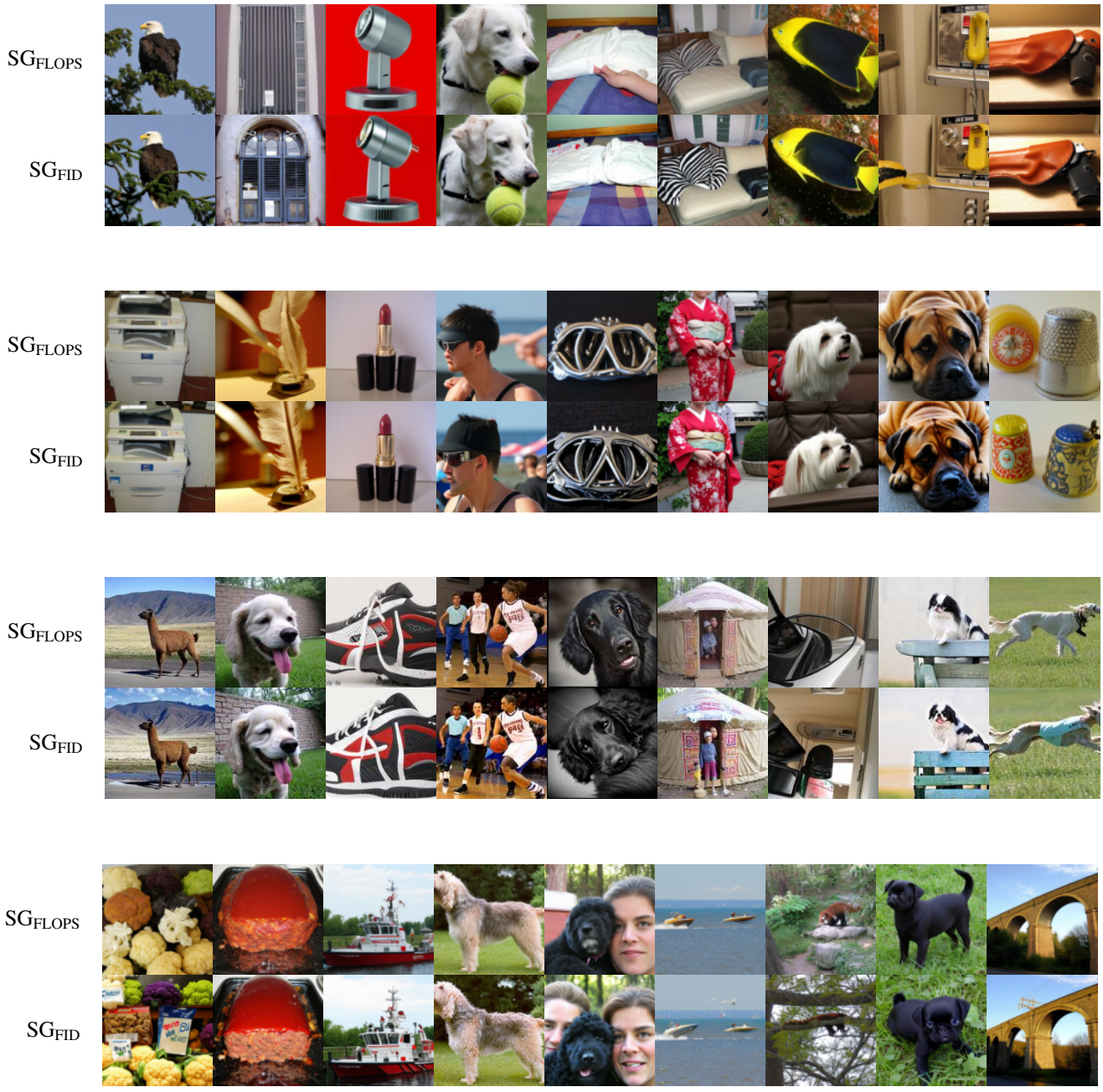

Figure A11. **Uncurated samples** of SG_{FLOPS} (top) and SG_{FID} (bottom) using $\omega = 2.5$.