Importance-Aware Pruning for Diffusion Models

Ba-Thinh Lam **Srijan Das** **Hieu Le**
Department of Machine Learning, UNC Charlotte, NC, USA
{blam5,sdas24,hle40}@charlotte.edu

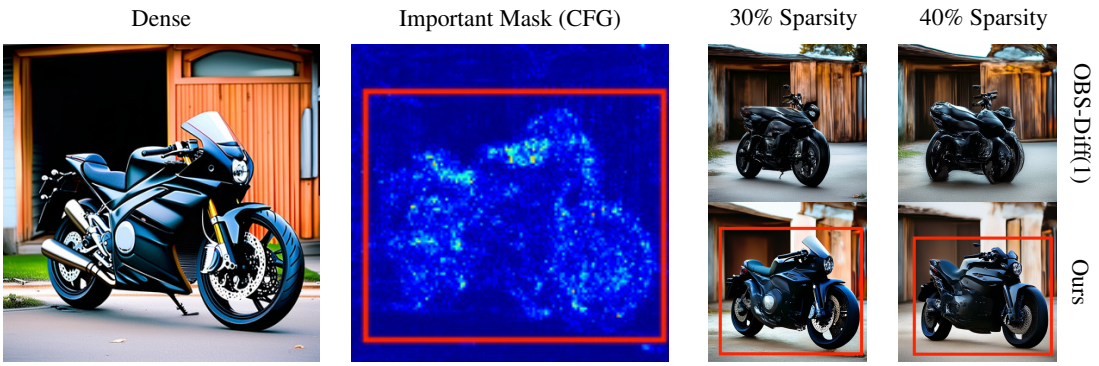


Figure 1: **Qualitative comparison on PixArt-Σ for Structured Pruning.** Our method (bottom) preserves subject integrity substantially better than the OBS-Diff baseline (top) by retaining semantically important regions, highlighted by red boxes. We achieve this by injecting importance signals such as classifier-free guidance (CFG) (2) into the pruning objective.

Abstract

We propose importance-aware pruning for diffusion models, a training-free framework that prioritizes preserving parameters critical to semantically salient image regions. To do so, we incorporate spatial importance maps- derived from conditioning signals or model attention- into the pruning objective. This produces parameter rankings aligned with perceptual relevance rather than uniform reconstruction error. On the MS-COCO dataset, our proposed approach consistently retains subject fidelity and structural correctness at high compression ratios where conventional pruning causes visible degradation. These results demonstrate that content-aware objectives are key to perceptually faithful compression of generative models.

1 Introduction

Pruning diffusion models (3; 4; 5; 6; 7; 8; 9; 10) is an important problem since diffusion models contain billions of parameters and require dozens of denoising steps at inference, making them expensive for many applications. Model pruning offers a promising solution by removing redundant parameters while preserving generation quality. Recent methods, such as OBS-Diff (1), have demonstrated that large-scale diffusion models can be pruned effectively without retraining.

Existing pruning methods typically treat all spatial locations equally. Yet not all regions of a generated image contribute equally to its perceived quality. Errors in prompt-relevant objects, faces, or foreground structures can significantly affect the perceived quality and semantic correctness of an image, while similar errors in background regions may be less noticeable (11; 12). This distinction becomes particularly important at high sparsity levels, where limited model capacity makes it difficult to preserve all visual details. As illustrated in Fig. 1, conventional pruning can preserve the overall

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

appearance of an image while degrading important subject details. This raises a fundamental question: Can model pruning prioritize the content that matters most for generation?

In this work, we propose importance-aware pruning, a training-free framework that incorporates spatial importance into parameter selection. To identify these regions, we construct spatial importance maps using signals such as classifier-free guidance (CFG) (2), Canny edge detection (13), or object detection (14). For example, CFG highlights regions strongly influenced by the text prompt, as illustrated in Fig. 1. These signals allow the pruning process to focus on different aspects of image content depending on the desired preservation goal. We then incorporate these importance maps into the pruning objective by assigning greater weight to reconstruction errors in important regions. For linear layers, this can be implemented by weighting the input activations used to estimate parameter importance. The framework therefore integrates naturally with second-order pruning methods such as OBS-Diff while retaining their parameter-selection and weight-compensation procedures. The resulting method remains one-shot and requires no retraining.

Our results show consistent improvements over previous methods, particularly at high sparsity levels, where our method better preserves subject integrity and structural details. We further demonstrate that different importance signals can prioritize different visual properties. In addition, category-targeted calibration allows pruning to favor specific semantic concepts, such as cats, airplanes, or people. More broadly, our work challenges the long-standing assumption that pruning should minimize average loss and instead argues that content-aware objectives are essential for perceptually faithful compression, connecting low-level parameter pruning with high-level semantic importance.

Contributions:

- We propose importance-aware pruning for diffusion models, grounded in the principle that parameters responsible for semantically critical image regions should be preferentially preserved during compression.
- We introduce a simple yet effective framework that supplements OBS saliency with spatial importance signals, and show that the signals can be flexibly derived from a variety of sources depending on the desired preservation goal.
- We show consistent improvements over the state-of-the-art method, OBS-Diff, across diffusion backbones and sparsity levels, with especially clear gains in subject fidelity and structural preservation under high compression ratios, for both structured and unstructured pruning.

2 Related Work

Diffusion Models. Diffusion models (15; 3; 16; 17) generate data by progressively transforming noise into structured samples through a learned reverse diffusion process. During training, clean data are gradually corrupted and the model learns to denoise across timesteps. This formulation has enabled state-of-the-art results across diverse generative tasks, including image synthesis (18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29), video generation (30; 31; 32; 33; 34; 35), 3D synthesis (36; 37; 38; 39; 40; 41; 42), and audio modeling (43; 44). Many practical systems follow the latent diffusion design (18), where images are encoded into latent tokens processed by a U-Net backbone augmented with attention and transformer layers. Extensions adapt this structure for multi-view (39) and video generation (32), while recent work replaces convolutional backbones with fully transformer-based diffusion architectures for improved scalability (21; 33). Classifier-free guidance (CFG) (2) improves conditional fidelity by combining conditional and unconditional predictions during sampling, and prior work suggests CFG implicitly highlights semantically important regions (45; 46). We build on this observation but use importance signals to guide parameter pruning rather than generation.

Diffusion Acceleration. Due to the high computational cost of diffusion inference, a wide range of efficiency strategies have been explored. Some methods require retraining, including improved architectures (18; 47; 48), structured pruning (49), compression approaches (50; 51; 52), step distillation (53; 54; 55; 56; 57), and consistency modeling (58; 59). Other approaches improve efficiency without retraining, such as advanced samplers (17; 60; 61; 62), feature caching (63; 64; 65; 66; 67; 68; 69; 70), quantization (71; 72; 73; 74; 75; 76), mixed-resolution inference (77), and token reduction (78; 79; 80). In particular, recent studies adapt token reduction

to diffusion inference (79; 81; 82; 83; 84; 78; 85; 86; 87; 88; 89), reducing computation by pruning or merging intermediate tokens during sampling. While effective for accelerating inference, these approaches operate on intermediate representations rather than model parameters and therefore do not reduce model size or memory footprint. Moreover, token selection is often heuristic or spatially uniform, which can degrade fidelity in visually important regions. In contrast, we study parameter-level pruning and explicitly bias pruning toward preserving semantically salient image content. Closest to our work is the importance-based token merging method of Wu *et al.* (90), which prioritizes less important tokens for merging based on CFG guidance. However, token merging is orthogonal to model pruning and, in practice, token merging can be applied on top of a pruned model for further acceleration.

Model Pruning. Pruning removes redundant parameters to reduce computation while maintaining model performance. Early methods rely on magnitude-based criteria (91) or first-order approximations, whereas second-order approaches such as Optimal Brain Damage (92) and Optimal Brain Surgeon (93) estimate parameter saliency using curvature information to better predict the effect of removal. Layer-wise formulations (94; 95) made second-order pruning tractable at scale, and subsequent methods such as SparseGPT (96) and Wanda (97) extended these ideas to billion-parameter language models. For diffusion models, existing approaches are either tailored to specific architectures (49; 98; 99; 100; 101; 50), dependent on costly retraining (49; 102; 103; 104; 50), or limited to structured sparsity (99; 101; 50; 105; 102; 106). OBS-Diff (1) is the first to apply second-order pruning to large-scale text-to-image diffusion models in a one-shot, training-free manner. However, its pruning objective treats all spatial locations uniformly, without regard for perceptual relevance. Our work extends this line by incorporating spatial importance signals into the saliency estimation, aligning pruning decisions with what matters most in the generated output.

3 Method

We propose importance-aware pruning, a framework that incorporates spatial importance signals into the OBS reconstruction objective. By weighting reconstruction errors according to spatial importance, the framework biases parameter selection toward preserving semantically or structurally important regions.

Preliminaries. For a layer l with weight matrix W_l and input activations $X_{l,t}$ at timestep t, the layer-wise pruning objective seeks a sparse weight matrix $\hat{W}_l$ that minimizes the reconstruction error:

$$\hat{W}_l^* = \arg\min_{\hat{W}_l} \mathbb{E}_{t\sim[1,T]} \left[\alpha_t \|\hat{W}_l X_{l,t} - W_l X_{l,t}\|_2^2\right] \quad \text{s.t.} \quad \text{Sparsity}(\hat{W}_l) = S, \tag{1}$$

where α_t is a timestep-dependent weight that places greater importance on earlier denoising steps, following the logarithmically decreasing schedule of (1).

3.1 OBS Pruning

The Optimal Brain Surgeon (OBS) framework solves the pruning objective by approximating the change in loss incurred by removing weight w_q via a second-order Taylor expansion, yielding a closed-form saliency score:

$$S_{l,q} = \frac{w_q^2}{2[H_l^{-1}]_{qq}}. \tag{2}$$

Parameters with smaller saliency are pruned first, with the remaining weights updated analytically to compensate for each removal via efficient Cholesky-based inverse Hessian updates (96).

However, this saliency estimate assumes that the model has converged to a minimum with negligible first-order gradient, which is rarely true in practice for diffusion models. As a result, the saliency scores are only approximate and may not faithfully reflect the true cost of removing each parameter. This motivates incorporating external importance guidance into the saliency estimation process. Specifically, we reweight the Hessian construction with a spatial importance signal, biasing saliency toward parameters that affect perceptually critical regions. This produces a pruning criterion that is better aligned with what the model should preserve, while keeping the structure and tractability of the OBS pipeline unchanged.

3.2 Spatial Importance Signal

Let $M_t \in \mathbb{R}^{H \times W}$ denote a spatial importance map at timestep t, with each entry reflecting the perceptual relevance of the corresponding spatial location. We construct this map from the classifier-free guidance (CFG) difference (2):

$$\Delta_t = \epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_\theta(\mathbf{x}_t, t, \varnothing), \tag{3}$$

where y denotes the text condition and $\varnothing$ denotes the unconditional input. The magnitude of this difference indicates how strongly the prompt influences the prediction at each location and serves as a proxy for spatial importance.

The raw CFG difference retains the latent channel dimension. We obtain a two-dimensional spatial map by averaging its absolute values across the C channels:

$$\tilde{M}_t(i,j) = \frac{1}{C} \sum_{k=1}^{C} \left| \Delta_{t,k,i,j} \right|. \tag{4}$$

We then normalize the map independently for each sample and timestep:

$$M_t = \frac{\tilde{M}_t - \min(\tilde{M}_t)}{\max(\tilde{M}_t) - \min(\tilde{M}_t) + \epsilon}, \tag{5}$$

where the minimum and maximum are taken over spatial locations and $\epsilon > 0$ ensures numerical stability.

Our framework is not restricted to CFG-based maps. Other signals that encode spatial relevance, such as Canny edge maps or object detection masks, can be used depending on the desired preservation goal, as demonstrated in Section 4.

3.3 Importance-Aware Hessian Construction

We incorporate the spatial importance signal into the pruning objective by weighting the reconstruction error at each spatial location by its importance:

$$\hat{W}_l^* = \arg\min_{\hat{W}_l} \sum_{t=1}^{T} \alpha_t \left\| M_t \odot \left(\hat{W}_l X_{l,t} - W_l X_{l,t} \right) \right\|_2^2 \quad \text{s.t.} \quad \text{Sparsity}(\hat{W}_l) = S. \tag{6}$$

For linear layers, since W_l acts along the channel dimension independently of the spatial dimension, it commutes with M_t, allowing us to define importance-filtered activations $X'_{l,t} = M_t \odot X_{l,t}$ and rewrite the objective as a standard OBS problem with $X_{l,t}$ replaced by $X'_{l,t}$. The importance-aware Hessian follows directly (see Appendix C for the full derivation):

$$H_{l,imp} = 2 \sum_{t=1}^{T} \alpha_t \mathbb{E}\left[X'_{l,t} X'^{\top}_{l,t} \right], \tag{7}$$

and substituting into the OBS saliency formula gives:

$$S_{l,q} = \frac{w_q^2}{2[H_{l,imp}^{-1}]_{qq}}. \tag{8}$$

Parameters whose activations concentrate in high-importance regions accumulate larger curvature and are preferentially retained; those whose influence is concentrated in perceptually irrelevant areas are preferentially removed. The entire modification reduces to a single substitution in the Hessian computation, with no change to the pruning pipeline or its computational complexity.

4 Experiments

4.1 Experimental Settings

Model and Dataset. We evaluate our method using Stable Diffusion 3 (SD3) Medium (5) and PixArt-Σ (107) as the base generative models. For the pruning calibration phase, we utilize 1,000 samples from the GCC3M dataset (108). To evaluate generative performance, we generate images using 1,000 prompts from the MS-COCO 2017 (109) validation set.

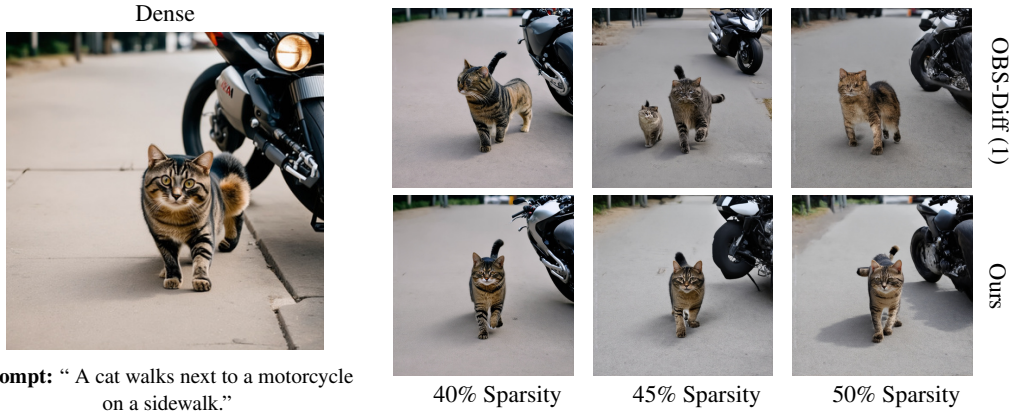
Baselines and Metrics. We compare our importance-aware framework against OBS-Diff (1), the current state-of-the-art second-order pruning method for diffusion models based on the Optimal Brain Surgeon framework (93). To evaluate generated image quality, we report CLIP Score (110) as a proxy for text alignment and ImageReward (111) as a human-aligned perceptual quality metric. For the category-targeted pruning experiment, we further assess semantic visual quality using MUSIQ (112).

Hyper-parameter setting. During the pruning phase, we set the classifier-free guidance (CFG) scale to 7.0 and 4.5 for SD3-Medium and PixArt-Σ, respectively, while setting the CFG scale to 7.0 for both during testing. We use 10 and 25 denoising steps for the pruning and testing phases, respectively, across all experiments. We choose a batch size of 2 during the pruning phase.

4.2 Quantitative Results

Table 1: Quantitative results of unstructured pruning of SD3-Medium on the GCC3M dataset (108).

Model	Sparsity	Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
SD3-Medium	Dense Model		32.26	0.98	72.68
	30%	OBS-Diff	**32.25**	0.96	**72.53**
		Ours (CFG)	32.24	**0.98**	72.41
	40%	OBS-Diff	32.30	0.92	71.02
		Ours (CFG)	**32.35**	**0.95**	**71.48**
	45%	OBS-Diff	32.25	0.84	68.76
		Ours (CFG)	**32.30**	**0.87**	**69.86**
	50%	OBS-Diff	32.16	0.71	65.98
		Ours (CFG)	**32.20**	**0.76**	**66.30**



Prompt: " A cat walks next to a motorcycle on a sidewalk."

Figure 2: **Qualitative comparison across different sparsity levels on SD3-Medium.** The top row shows OBS-Diff, which exhibits artifacts and loss of subject coherence as sparsity increases. The bottom row shows our Importance-Aware OBS, which better preserves the structural and semantic integrity of the generated image by prioritizing critical feature regions.

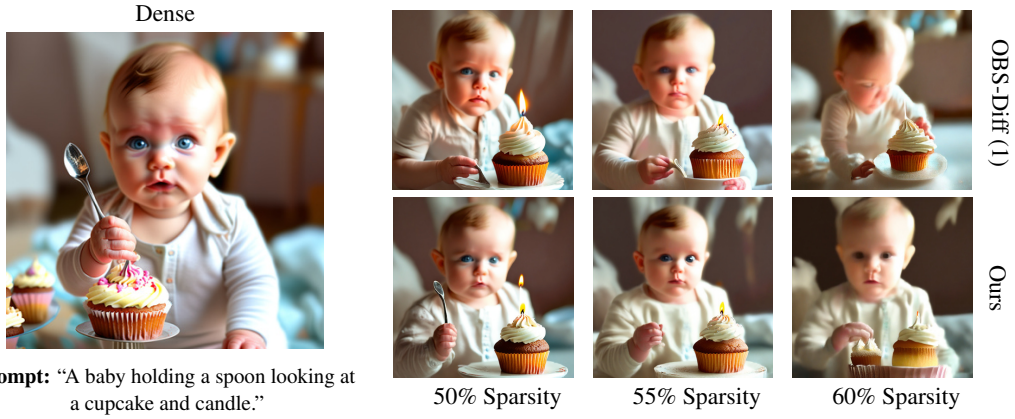
SD3-Medium. Table 1 compares our approach to OBS-Diff on SD3-Medium and shows that injecting the CFG mask during pruning outperforms the baseline, particularly as sparsity levels increase. By leveraging semantic guidance from the CFG map, our method maintains higher alignment (CLIP Score) and human preference (ImageReward), preserving the model's generative quality even at 50% sparsity, where standard OBS-Diff begins to exhibit structural degradation. As shown in Fig. 2, our approach consistently preserves the object integrity and fine-grained semantic textures across various sparsity levels, even at an aggressive level of 60%.

PixArt-Σ. As shown in Table 2, our approach to pruning with the CFG mask outperforms the baseline OBS-Diff on PixArt-Σ across several aggressive sparsity levels. While both methods maintain high CLIP Scores, our method demonstrates superior resilience in human preference metrics (ImageReward). Notably, at 60% sparsity, our approach achieves an ImageReward of 0.52 compared to 0.49 for OBS-Diff. Fig. 3 further shows the accurately important semantic preservation of our

approach relative to OBS-Diff. This indicates that by focusing the Hessian approximation on regions of high classifier-free guidance, we effectively prioritize the preservation of weights responsible for the primary semantic subjects of the prompt.

Table 2: Quantitative comparison between OBS-Diff and Ours (CFG) on PixArt-Σ using 1000 samples from GCC3M for unstructured pruning.

Model	Sparsity	Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
PixArt-Σ	Dense Model		31.86	0.94	71.11
	45%	OBS-Diff	31.63	0.78	70.12
		Ours (CFG)	**31.73**	**0.85**	**70.77**
	50%	OBS-Diff	31.62	0.76	70.07
		Ours (CFG)	**31.71**	**0.78**	**70.34**
	55%	OBS-Diff	31.65	0.64	68.25
		Ours (CFG)	**31.69**	**0.70**	**69.25**
	60%	OBS-Diff	31.44	0.49	65.24
		Ours (CFG)	**31.54**	**0.52**	**67.13**



Prompt: "A baby holding a spoon looking at a cupcake and candle."

Figure 3: **Qualitative comparison across different sparsity levels on PixArt-Σ for Unstructured Pruning.** The top row shows OBS-Diff, which exhibits artifacts and loss of subject coherence as sparsity increases. The bottom row shows our Importance-Aware OBS, which better preserves the structural and semantic integrity of the generated image by prioritizing critical feature regions.

4.3 Flexibility to Different Importance Guidance Signals

Our framework can use different importance signals without changing the pruning formulation. Besides the CFG mask, we evaluate Canny edge masks and object-local CFG masks. The Canny mask highlights image edges, while the object-local CFG mask is obtained by detecting objects with YOLOv8 (14) and retaining the corresponding regions from the original CFG map.

Tab. 3 demonstrates the flexibility of our importance-aware pruning framework. The CFG signal can be replaced by a simple, model-agnostic cue such as Canny edges, making the approach readily applicable to diffusion models where CFG-based maps are unavailable or inconvenient to extract. With Canny guidance, the pruning objective shifts toward preserving finer structural details and local boundaries. Moreover, by incorporating an object detector on top of the CFG, our framework can be steered toward object-centric preservation. This customization yields consistent metric improvements and better maintains object structures at high sparsity.

Fig. 4 confirms the same trend qualitatively. Object-local CFG better preserves the motorcycle parts (e.g., windshield and kickstand) by focusing on detected object regions, while Canny produces sharper wheel textures by emphasizing edge structure. These results show that our method is modular: different guidance signals can be plugged in to preserve different semantic or structural priors under pruning.

Table 3: Quantitative comparison including Ours (Canny) and Ours (CFG+Detector) on PixArt-Σ for Unstructured Pruning. Best results per sparsity level are bolded.

Model	Sparsity	Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
PixArt-Σ	**Dense Model**		31.86	0.94	71.11
	45%	OBS-Diff	31.63	0.78	70.12
		Ours (CFG)	**31.73**	**0.85**	70.77
		Ours (CFG+Detector)	**31.73**	0.82	**70.87**
		Ours (Canny)	**31.73**	0.82	70.81
	50%	OBS-Diff	31.62	0.76	70.07
		Ours (CFG)	31.71	0.78	70.34
		Ours (CFG+Detector)	31.71	0.78	**70.59**
		Ours (Canny)	**31.75**	**0.81**	70.34
	55%	OBS-Diff	31.65	0.64	68.25
		Ours (CFG)	31.69	0.70	69.25
		Ours (CFG+Detector)	31.79	0.75	**69.62**
		Ours (Canny)	**31.83**	**0.76**	69.12
	60%	OBS-Diff	31.44	0.49	65.24
		Ours (CFG)	31.54	0.52	67.13
		Ours (CFG+Detector)	**31.60**	0.50	**67.27**
		Ours (Canny)	31.52	**0.53**	65.91

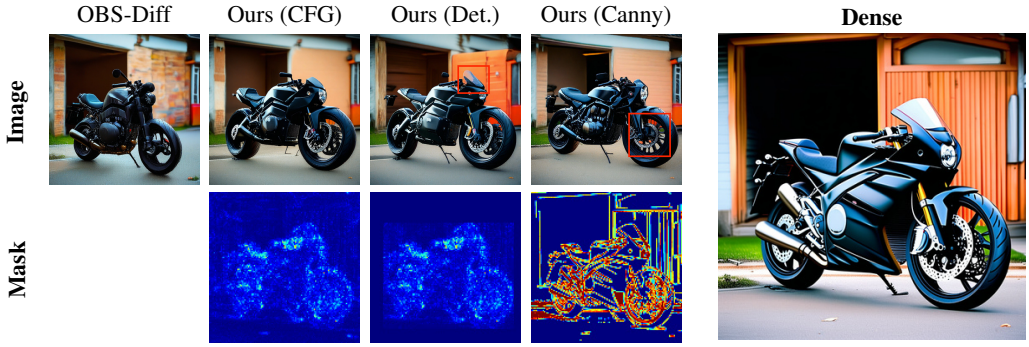


Figure 4: **Pruned Models with Different Important Signals.** Comparison of the OBS-Diff baseline against our importance-aware variants (CFG, CFG+Detector, and Canny). The second row visualizes the corresponding saliency masks. **Prompt:** "black Honda motorcycle parked in front of a garage."

4.4 Biasing Pruning Toward Targeted Categories

Our framework suggests a broader view of pruning: compression can be biased toward preserving user-specified properties rather than treating all visual content uniformly. We study category preservation as a simple instance of this idea. A straightforward way to introduce this bias is through the calibration data: We construct category-targeted pruning subsets whose prompts contain a predefined category, such as “airplane”, “cat”, and “woman”. We compare OBS-Diff, our general model pruned on the original category-diverse calibration set, and our targeted model pruned on the corresponding category-specific subset. All models are evaluated on the same category-targeted test prompts.

Tab. 4 shows that category-specific calibration already steers pruning toward the target concept. At 60% sparsity, Ours (target) outperforms OBS-Diff and our general variant on most quality metrics. For example, it improves MUSIQ from 69.52 to 71.64 on Cat and achieves the best MUSIQ score of 67.40 on Woman. These results support our hypothesis that targeted calibration shifts the Hessian approximation and saliency estimates toward parameters important for the selected category. Additional spatial guidance further strengthens this bias, making targeted preservation more effective under high sparsity. Fig.5 shows two visual examples for two categories - “women” and “cat”. For more qualitative results, please refer to Appendix B.4.

Table 4: Quantitative results of unstructured category-targeted pruning on PixArt-Σ at 60% sparsity.

Category	Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
Woman	OBS-Diff (general)	32.36	0.72	65.08
	OBS-Diff (target)	32.25	0.75	65.75
	Ours (general)	32.31	0.76	66.74
	Ours (target)	**32.38**	**0.85**	**67.40**
Cat	OBS-Diff (general)	32.58	0.28	64.35
	OBS-Diff (target)	32.45	0.29	69.81
	Ours (general)	**32.66**	**0.41**	69.52
	Ours (target)	32.28	0.34	**71.64**
Airplane	OBS-Diff (general)	**29.35**	0.45	67.81
	OBS-Diff (target)	28.62	0.43	69.90
	Ours (general)	29.22	**0.62**	68.19
	Ours (target)	28.76	0.51	**70.81**



Figure 5: **Target-category pruning at 60% sparsity.** Category-targeted pruning improves preservation of the specified category, and our method further improves target fidelity and visual quality.

4.5 Structured Pruning

Our formulation can be directly extended to structured pruning by following the aggregation strategy in OBS-Diff. Table 5 shows that our method consistently outperforms OBS-Diff across sparsity levels in this setting. At 40% sparsity, OBS-Diff drops to an ImageReward of -0.04, while our CFG-guided pruning achieves 0.30, indicating better preservation of semantic and structural details. This is also reflected in fig. 1 and fig. 6, with additional qualitative results in Appendix B.5.

Table 5: Quantitative comparison between OBS-Diff and Ours (CFG) on PixArt-Σ using 1000 samples from GCC3M for structured pruning.

Model	Sparsity	Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
PixArt-Σ	Dense Model		31.86	0.94	71.11
	20%	OBS-Diff	31.65	0.74	69.57
		Ours (CFG)	**31.68**	**0.79**	**71.37**
	30%	OBS-Diff	**31.36**	0.38	66.98
		Ours (CFG)	31.30	**0.54**	**69.85**
	40%	OBS-Diff	30.84	-0.04	65.17
		Ours (CFG)	**31.08**	**0.30**	**68.64**



Figure 6: **Structured Pruning on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** An Egyptian airlines plane landing at an airport.

4.6 Ablation Study

We conduct an ablation study on PixArt-Σ to analyze the impact of the Classifier-Free Guidance (CFG) scale during the pruning phase.

The impact of the guidance scale. We analyze the sensitivity of mask extraction to the Classifier-Free Guidance (CFG) scale used during the pruning phase. Specifically, we compare pruning PixArt-Σ under guidance scales of GS $= 4.5$ and GS $= 7.0$, while keeping the evaluation guidance scale fixed at 7.0. As shown in Table 6, our method consistently outperforms OBS-Diff across both 45% and 50% sparsity levels under both guidance configurations.

While OBS-Diff exhibits performance degradation when reducing the pruning CFG scale from 7.0 to 4.5 (e.g., CLIP Score drops from 31.71 to 31.63 at 45% sparsity), our approach maintains robust text alignment (CLIP Score ≥ 31.71) across all settings. Furthermore, at GS $= 4.5$, our method achieves its highest ImageReward of 0.85 at 45% sparsity and 0.78 at 50% sparsity. This demonstrates that our importance-guided saliency estimation effectively preserves critical parameter weights without relying on high guidance amplification during mask extraction.

Table 6: Quantitative results of unstructured pruning of PixArt-Σ at 45% and 50% sparsity levels, comparing pruning guidance scales of 4.5 and 7.0 (tested at 7.0). The best results per guidance scale are highlighted in bold.

		Guidance Scale = 4.5		**Guidance Scale = 7.0**	
Sparsity	**Method**	**CLIP Score ↑**	**ImageReward ↑**	**CLIP Score ↑**	**ImageReward ↑**
Dense Model		31.86	0.94	31.86	0.94
45%	OBS-Diff	31.63	0.78	31.71	0.79
	Ours	**31.73**	**0.85**	**31.72**	**0.82**
50%	OBS-Diff	31.62	0.76	31.70	0.76
	Ours	**31.71**	**0.78**	**31.73**	**0.77**

4.7 Human Evaluation

To further assess perceptual quality, we conduct a user study in which participants rate generated images on a five-point scale from 1 (collapsed) to 5 (excellent). As shown in Tab. 7, our method is strongly preferred over OBS-Diff, achieving a substantially higher human score of 3.97 compared to 2.59 at 60% sparsity. Notably, despite this clear improvement in human-perceived quality, our method receives a slightly lower CLIPScore. This discrepancy further highlights that automatic metrics do not always reliably reflect perceptual image quality. We provide additional analysis and qualitative examples in Appendix A.

Table 7: Quantitative Comparison on 20-Sample Evaluation Set on PixArt-Σ at 60% Sparsity. Bold values indicate superior performance in automated metrics.

Method	CLIP Score ↑	ImageReward ↑	MUSIQ ↑	Human Score ↑
OBS-Diff	**31.90**	0.72	69.83	2.59
Ours	31.06	**0.78**	**72.93**	**3.97**

5 Limitations

Our framework inherits the limitations of the chosen importance signal. CFG-based masks are training-free, but they can be noisy or incomplete for abstract prompts, small objects, and heavily cluttered scenes. In addition, our method is designed as a post-training pruning approach and does not explore whether finetuning can recover further quality at very high sparsity. For unstructured pruning, practical latency gains still require hardware or sparse-kernel support. Finally, our evaluation is centered on text-to-image generation; extending the same principle to video diffusion, editing, personalization, and other conditional generation tasks remains future work.

6 Conclusion

We introduced an importance-aware pruning framework for diffusion models that shifts compression from uniformly preserving all reconstruction errors to preserving the errors that matter most for generation. By incorporating semantic importance signals into second-order pruning, our method better protects prompt-relevant regions and fine-grained structures without requiring retraining or architectural changes. Across unstructured, structured, and category-targeted pruning settings, our approach consistently improves semantic preservation over OBS-Diff, especially under aggressive sparsity where baseline pruning often collapses subject structure. These results show that diffusion pruning should not be treated as a purely parameter-level compression problem: effective compression must account for where and what the model needs to preserve. We believe this work supports a broader view of guided model compression, where pruning objectives can be biased toward user-specified semantic, spatial, or task-level priorities.

Acknowledgments This work was supported by faculty startup funds from the College of Computing and Informatics (CCI) at the University of North Carolina at Charlotte.

References

[1] J. Zhu, H. Wang, M. Su, Z. Wang, and H. Wang, "OBS-diff: Accurate pruning for diffusion models in one-shot," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=eYdPMpW5os

[2] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023.

[5] 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," in *Forty-first international conference on machine learning*, 2024.

[6] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao *et al.*, "Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding," *arXiv preprint arXiv:2405.08748*, 2024.

[7] H. Li, S. Lal, Z. Li, Y. Xie, Y. Wang, Y. Zou, O. Majumder, R. Manmatha, Z. Tu, S. Ermon *et al.*, "Efficient scaling of diffusion transformers for text-to-image generation," *arXiv preprint arXiv:2412.12391*, 2024.

[8] B. Tang, B. Zheng, S. Paul, and S. Xie, "Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 28 586–28 595.

[9] E. Xie, J. Chen, Y. Zhao, J. Yu, L. Zhu, C. Wu, Y. Lin, Z. Zhang, M. Li, J. Chen *et al.*, "Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer," *arXiv preprint arXiv:2501.18427*, 2025.

[10] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, K. Lacey, Y. Levi, C. Li, D. Lorenz, J. Müller, D. Podell, R. Rombach, H. Saini, A. Sauer, and L. Smith, "Flux.1 kontext: Flow matching for in-context image generation and editing in latent space," 2025. [Online]. Available: https://arxiv.org/abs/2506.15742

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:3108956

[12] R. Xue, J. Xu, S. Mondal, H. Le, G. J. Zelinsky, M. Hoai, and D. Samaras, "Few-shot personalized scanpath prediction," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 497–13 507, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277627903

[13] J. F. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, pp. 679–698, 1986. [Online]. Available: https://api.semanticscholar.org/CorpusID:13284142

[14] R. Varghese and S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6.

[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[19] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[20] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[21] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023.

[22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248097655

[23] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[24] A. Graikos, S. Yellapragada, M.-Q. Le, S. Kapse, P. Prasanna, J. Saltz, and D. Samaras, "Learned representation-guided diffusion models for large-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8532–8542.

[25] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[26] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo, "Raphael: Text-to-image generation via large mixture of diffusion paths," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[28] T. Karras, M. Aittala, T. Kynkäänniemi, J. Lehtinen, T. Aila, and S. Laine, "Guiding a diffusion model with a bad version of itself," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52 996–53 021, 2025.

[29] J. Xu, H. Le, and D. Samaras, "Assessing sample quality via the latent space of generative models," in *ECCV*, 2024.

[30] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[31] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[32] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[33] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," March 2024. [Online]. Available: https://github.com/hpcaitech/Open-Sora

[34] P.-Y. Lab and T. A. etc., "Open-sora-plan," Apr. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10948109

[35] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman *et al.*, "Video generation models as world simulators. 2024," *URL https://openai.com/research/video-generation-models-as-world-simulators*, vol. 3, 2024.

[36] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[37] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 9298–9309.

[38] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9970–9980.

[39] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *arXiv preprint arXiv:2310.15110*, 2023.

[40] G. Li, H. Zheng, C. Wang, C. Li, C. Zheng, and D. Tao, "3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2211.14108*, 2022.

[41] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2837–2845.

[42] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," *arXiv preprint arXiv:2212.08751*, 2022.

[43] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[44] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[45] J. Zhao, H. Zheng, C. Wang, L. Lan, and W. Yang, "Magicfusion: Boosting text-to-image generation performance by fusing diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 592–22 602.

[46] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "High-fidelity person-centric subject-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7675–7684.

[47] P. Pernias, D. Rampas, M. L. Richter, C. J. Pal, and M. Aubreville, "Würstchen: An efficient architecture for large-scale text-to-image diffusion models," *arXiv preprint arXiv:2306.00637*, 2023.

[48] B.-K. Kim, H.-K. Song, T. Castells, and S. Choi, "On architectural compression of text-to-image diffusion models," 2023.

[49] G. Fang, X. Ma, and X. Wang, "Structural pruning for diffusion models," in *Advances in Neural Information Processing Systems*, 2023.

[50] Y. Zhao, Y. Xu, Z. Xiao, and T. Hou, "Mobilediffusion: Subsecond text-to-image generation on mobile devices," *arXiv preprint arXiv:2311.16567*, 2023.

[51] X. Yang, D. Zhou, J. Feng, and X. Wang, "Diffusion probabilistic model made slim," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2023, pp. 22 552–22 562.

[52] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren, "Snapfusion: Text-to-image diffusion model on mobile devices within two seconds," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[53] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[54] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 297–14 306.

[55] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2025, pp. 87–103.

[56] X. Liu, X. Zhang, J. Ma, J. Peng *et al.*, "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation," in *The Twelfth International Conference on Learning Representations*, 2023.

[57] A. Habibian, A. Ghodrati, N. Fathima, G. Sautiere, R. Garrepalli, F. Porikli, and J. Petersen, "Clockwork diffusion: Efficient generation with model-step distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8352–8361.

[58] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[59] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

[60] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," *arXiv preprint arXiv:2202.09778*, 2022.

[61] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[62] ——, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[63] X. Ma, G. Fang, and X. Wang, "Deepcache: Accelerating diffusion models for free," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 762–15 772.

[64] F. Wimbauer, B. Wu, E. Schoenfeld, X. Dai, J. Hou, Z. He, A. Sanakoyeu, P. Zhang, S. Tsai, J. Kohler *et al.*, "Cache me if you can: Accelerating diffusion models through block caching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6211–6220.

[65] P. Chen, M. Shen, P. Ye, J. Cao, C. Tu, C.-S. Bouganis, Y. Zhao, and T. Chen, "Delta-dit: A training-free acceleration method tailored for diffusion transformers," *arXiv preprint arXiv:2406.01125*, 2024.

[66] X. Zhao, X. Jin, K. Wang, and Y. You, "Real-time video generation with pyramid attention broadcast," *arXiv preprint arXiv:2408.12588*, 2024.

[67] K. Kahatapitiya, H. Liu, S. He, D. Liu, M. Jia, M. S. Ryoo, and T. Xie, "Adaptive caching for faster video generation with diffusion transformers," *arXiv preprint arXiv:2411.02397*, 2024.

[68] J. So, J. Lee, and E. Park, "Frdiff: Feature reuse for universal training-free acceleration of diffusion models," *arXiv preprint arXiv:2312.03517*, 2023.

[69] Z. Lv, C. Si, J. Song, Z. Yang, Y. Qiao, Z. Liu, and K.-Y. K. Wong, "Fastercache: Training-free video diffusion model acceleration with high quality," *arXiv preprint arXiv:2410.19355*, 2024.

[70] S. Li, T. Hu, F. S. Khan, L. Li, S. Yang, Y. Wang, M.-M. Cheng, and J. Yang, "Faster diffusion: Rethinking the role of unet encoder in diffusion models," *CoRR*, 2023.

[71] X. Li, Y. Liu, L. Lian, H. Yang, Z. Dong, D. Kang, S. Zhang, and K. Keutzer, "Q-diffusion: Quantizing diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 535–17 545.

[72] L. Chen, Y. Meng, C. Tang, X. Ma, J. Jiang, X. Wang, Z. Wang, and W. Zhu, "Q-dit: Accurate post-training quantization for diffusion transformers," *arXiv preprint arXiv:2406.17343*, 2024.

[73] Y. He, L. Liu, J. Liu, W. Wu, H. Zhou, and B. Zhuang, "Ptqd: Accurate post-training quantization for diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[74] H. Wang, Y. Shang, Z. Yuan, J. Wu, J. Yan, and Y. Yan, "Quest: Low-bit diffusion model quantization via efficient selective finetuning," *arXiv preprint arXiv:2402.03666*, 2024.

[75] J. Deng, S. Li, Z. Wang, H. Gu, K. Xu, and K. Huang, "Vq4dit: Efficient post-training vector quantization for diffusion transformers," *arXiv preprint arXiv:2408.17131*, 2024.

[76] J. So, J. Lee, D. Ahn, H. Kim, and E. Park, "Temporal dynamic quantization for diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[77] H. Wu, J. Xu, Q. Miao, D. Samaras, and H. Le, "Phase-aligned rope for mixed-resolution diffusion transformer," in *ECCV*. Springer, 2026, pp. 272–293.

[78] K. Kahatapitiya, A. Karjauv, D. Abati, F. Porikli, Y. M. Asano, and A. Habibian, "Object-centric diffusion for efficient video editing," in *European Conference on Computer Vision*. Springer, 2025, pp. 91–108.

[79] D. Bolya and J. Hoffman, "Token merging for fast stable diffusion," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4599–4603, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257833518

[80] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster," *ArXiv*, vol. abs/2210.09461, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252968113

[81] M. Kim, S. Gao, Y.-C. Hsu, Y. Shen, and H. Jin, "Token fusion: Bridging the gap between token pruning and token merging," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1383–1392.

[82] J. B. Haurum, S. Escalera, G. W. Taylor, and T. B. Moeslund, "Agglomerative token clustering," in *European Conference on Computer Vision*. Springer, 2025, pp. 200–218.

[83] H. Wang, D. Liu, Y. Kang, Y. Li, Z. Lin, N. K. Jha, and Y. Liu, "Attention-driven training-free efficiency enhancement of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 080–16 089.

[84] X. Li, C. Ma, X. Yang, and M.-H. Yang, "Vidtome: Video token merging for zero-shot video editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7486–7495.

[85] W. Zhao, Y. Han, J. Tang, K. Wang, Y. Song, G. Huang, F. Wang, and Y. You, "Dynamic diffusion transformer," *arXiv preprint arXiv:2410.03456*, 2024.

[86] F. Liang, A. Kodaira, C. Xu, M. Tomizuka, K. Keutzer, and D. Marculescu, "Looking backward: Streaming video-to-video translation with feature banks," *arXiv preprint arXiv:2405.15757*, 2024.

[87] K. Wang, J. Chen, H. Li, Z. Mi, and J. Zhu, "Sparsedm: Toward sparse efficient diffusion models," *arXiv preprint arXiv:2404.10445*, 2024.

[88] E. Smith, N. Saxena, and A. Saha, "Todo: Token downsampling for efficient generation of high-resolution images," *arXiv preprint arXiv:2402.13573*, 2024.

[89] W. Lu, S. Zheng, Y. Xia, and S. Wang, "Toma: Token merging with attention for diffusion models," 2024.

[90] H. Wu, J. Xu, H. Le, and D. Samaras, "Importance-based token merging for efficient image and video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 4983–4995.

[91] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems*, 2015.

[92] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, 1989.

[93] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 1992.

[94] X. Dong, S. Chen, and S. Pan, "Learning to prune deep neural networks via layer-wise optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 2017.

[95] E. Frantar and D. Alistarh, "Optimal brain compression: A framework for accurate post-training quantization and pruning," in *Advances in Neural Information Processing Systems*, 2022.

[96] ——, "SparseGPT: Massive language models can be accurately pruned in one shot," in *International Conference on Machine Learning*, 2023.

[97] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," in *International Conference on Learning Representations*, 2024.

[98] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren, "SnapFusion: Text-to-image diffusion model on mobile devices within two seconds," in *Advances in Neural Information Processing Systems*, 2023.

[99] B.-K. Kim, H.-K. Song, T. Castells, and S. Choi, "BK-SDM: A lightweight, fast, and cheap version of stable diffusion," in *European Conference on Computer Vision*, 2024.

[100] T. Castells, H.-K. Song, B.-K. Kim, and S. Choi, "Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 821–830.

[101] D. Zhang, S. Li, C. Chen, Q. Xie, and H. Lu, "Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models," *arXiv preprint arXiv:2404.11098*, 2024.

[102] H. Zhu, D. Tang, J. Liu, M. Lu, J. Zheng, J. Peng, D. Li, Y. Wang, F. Jiang, L. Tian *et al.*, "Dip-go: A diffusion pruner via few-step gradient optimization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 92 581–92 604, 2024.

[103] G. Fang, K. Li, X. Ma, and X. Wang, "Tinyfusion: Diffusion transformers learned shallow," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 144–18 154.

[104] Y. Zhang, E. Jin, Y. Dong, A. Khakzar, P. Torr, J. Stegmaier, and K. Kawaguchi, "Effortless efficiency: Low-cost pruning of diffusion models," *arXiv preprint arXiv:2412.02852*, 2024.

[105] A. Ganjdanesh, R. Shirkavand, S. Gao, and H. Huang, "Not all prompts are made equal: Prompt-based pruning of text-to-image diffusion models," *arXiv preprint arXiv:2406.12042*, 2024.

[106] B. Chen, G. Li, R. Wu, X. Zhang, J. Chen, J. Zhang, and L. Zhang, "Adversarial diffusion compression for real-world image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025, pp. 28 208–28 220.

[107] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 74–91.

[108] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[109] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[110] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[111] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15 903–15 935, 2023.

[112] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5148–5157.

[113] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan, "Evaluating text-to-visual generation with image-to-text generation," *arXiv preprint arXiv:2404.01291*, 2024.

[114] D. Ghosh, H. Hajishirzi, and L. Schmidt, "Geneval: An object-focused framework for evaluating text-to-image alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 132–52 152, 2023.

A Evaluation metrics and User Study.

Evaluating compressed text-to-image models remains challenging because automatic metrics only partially capture the failures introduced by pruning. CLIP-based metrics are useful for measuring coarse image-text alignment, but they are often insensitive to localized semantic damage, object deformation, missing fine details, and spatial relation errors. This limitation is consistent with recent text-to-image evaluation studies showing that CLIPScore often behaves as a global alignment measure and can fail on compositional prompts involving objects, attributes, and relations (113; 114). This weakness is especially relevant in our setting: pruning does not always destroy the entire image, but often damages foreground subjects, object boundaries, limbs, faces, or other semantically important regions. Such failures can still receive reasonable CLIP scores as long as the image remains broadly related to the prompt.

We therefore include ImageReward and MUSIQ as complementary metrics. ImageReward provides a stronger preference-oriented signal because it is trained from human preference annotations and is designed to better reflect perceptual quality and prompt fidelity in text-to-image generation (111). MUSIQ, on the other hand, measures image quality without relying directly on text-image alignment, making it useful for detecting visual degradation such as blur, artifacts, and unnatural structure. In our benchmark results, ImageReward and MUSIQ generally separate visually coherent generations from degraded ones more clearly than CLIPScore, especially at high sparsity. However, both remain automatic metrics and should not be treated as complete substitutes for human judgment.

For this reason, we further conduct a user study to directly measure perceived image quality. Each participant rates every generated image on a five-level quality scale: 1-*Collapsed* for incoherent, broken, or unusable images; 2-*Distorted* for images with major defects and poor visual quality; 3-*Degraded* for images with visible artifacts but still recognizable content; 4-*Good* for mostly natural images with minor artifacts; and 5-*Excellent* for clean, natural, artifact-free images. This protocol allows us to compare automatic metric preferences against human judgments at the individual-sample level.

Our results in Tab.8 show that ImageReward and MUSIQ align better with human ratings than CLIPScore, but none of the automatic metrics is perfect. Qualitative examples reveal cases where CLIPScore remains similar, or even favors the baseline, despite clear degradation in subject integrity, edge structure, or object-level detail, as can be seen in Fig.7. These observations support our main evaluation choice: the benefit of importance-aware pruning should not be judged solely by global embedding-based metrics. Its advantage is most visible in preserving the semantically important regions that humans notice first, but that CLIP-style metrics may under-penalize.



OBS-Diff



Ours (CFG)

Method	Human Score ↑	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
OBS-Diff	1.65	**29.71**	1.34	73.58
Ours	**3.59**	28.79	**1.60**	**74.46**

(a) Sample A (Index: 17). Human evaluators strongly prefer our method (3.59 vs 1.65). ImageReward and MUSIQ correctly align with the human preference by scoring our method higher, whereas the CLIP Score incorrectly favors the baseline. **Prompt:** "An airplane high in the sky over some electrical wires."



OBS-Diff



Ours (CFG)

Method	Human Score ↑	CLIP Score ↑	ImageReward ↑	MUSIQ ↑
OBS-Diff	2.29	**29.30**	1.71	65.07
Ours	**3.59**	28.34	**1.93**	**70.41**

(b) Sample B (Index: 7). Human evaluators again show a strong preference for our method (3.59 vs 2.29). While ImageReward and MUSIQ successfully align with human perception, the CLIP Score incorrectly assigns a higher value to the baseline. **Prompt:** "A woman is riding her bike down the street in front of traffic".

Figure 7: Visual comparisons highlighting the misalignment between automated metrics and human perception at 60% sparsity. Bold values indicate the higher score. These examples demonstrate the limitations of CLIP Score in reliably capturing human-perceived visual quality following pruning, a gap bridged more effectively by ImageReward and MUSIQ.

Table 8: Detailed per-sample alignment between mean human scores and automated metric preferences. ✓indicates the metric preference matches the human preference ($H_O > H_B$), while ✗indicates the metric preference contradicts the human preference.

Index	Human (B)	Human (O)	CLIP Score	ImageReward	MUSIQ
1	2.00	**3.59**	✗	✗	✓
2	2.53	**4.06**	✓	✗	✓
3	2.59	**3.29**	✗	✗	✓
4	2.41	**3.18**	✗	✗	✓
5	2.24	**3.76**	✓	✓	✗
6	2.18	**4.00**	✓	✗	✓
7	2.29	**3.59**	✗	✓	✓
8	2.53	**3.59**	✗	✓	✓
9	3.00	**4.59**	✗	✗	✓
10	3.35	**4.18**	✗	✗	✗
11	2.88	**4.59**	✗	✗	✗
12	3.00	**4.53**	✗	✓	✓
13	3.18	**4.56**	✗	✓	✓
14	3.47	**4.65**	✓	✓	✗
15	2.65	**4.76**	✗	✗	✓
16	2.94	**4.24**	✗	✓	✓
17	1.65	**3.59**	✗	✓	✓
18	2.53	**3.76**	✗	✓	✗
19	2.12	**3.59**	✗	✓	✓
20	2.18	**3.29**	✓	✗	✗
Average Alignment With Human Score			25.00%	50.00%	70.00%

B Qualitative Results

B.1 Results on PixArt-Σ

Fig. 8 - Fig. 17 illustrate qualitative comparison across aggressively increasing sparsity using PixArt-Σ with various prompts. We compare our approach using the original CFG mask as an important guidance signal to the OBS-Diff baseline.

B.2 Results on SD3-Medium

Fig. 18 - Fig. 22 illustrate qualitative comparison across aggressively increasing sparsity using SD3-Medium with various prompts. We compare our approach using the original CFG mask as an important guidance signal to the OBS-Diff baseline.

B.3 Flexibility Across Different Importance Signals

Fig. 37 - Fig. 41 illustrate comparison of the baseline OBS-Diff against our importance-aware variants guided by different signals (CFG, Object Detector, Canny Edge) at 45%, 50%, 55%, and 60% unstructured sparsity.

B.4 Biasing Pruning Toward Targeted Categories

Woman Fig. 23 shows a qualitative comparison among Ours (Target), Ours (General), and OBS-Diff (Target) at 60% sparsity level for woman-targeted pruning.

Cat Fig. 24 shows a qualitative comparison among Ours (Target), Ours (General), and OBS-Diff (Target) at 60% sparsity level for cat-targeted pruning.

Motorcycle Fig. 25 shows a qualitative comparison among Ours (Target), Ours (General), and OBS-Diff (Target) at 60% sparsity level for motorcycle-targeted pruning.

Airplane Fig. 26 shows a qualitative comparison among Ours (Target), Ours (General), and OBS-Diff (Target) at 60% sparsity level for airplane-targeted pruning.

B.5 Structured Pruning

Fig. 27 - Fig. 36 illustrate qualitative comparison across aggressively increasing sparsity using PixArt-Σ with various prompts for Structured Pruning. We compare our approach using the original CFG mask as an important guidance signal to the OBS-Diff baseline.

C Derivation of Importance-Aware OBS Saliency

We derive the closed-form saliency score used by our importance-aware OBS formulation. For layer l, let $X'_{l,t} = M_t \odot X_{l,t}$ denote the importance-filtered activation at timestep t. The layer-wise reconstruction objective is

$$\mathcal{J}(\hat{W}_l) = \sum_{t=1}^{T} \alpha_t \left\| \hat{W}_l X'_{l,t} - W_l X'_{l,t} \right\|_F^2 . \tag{9}$$

This objective is minimized at $\hat{W}_l = W_l$, where the reconstruction error and the first-order gradient are both zero. Therefore, the second-order expansion around W_l is exact for this quadratic surrogate objective.

Following standard OBS-style layer-wise pruning, we derive the update for one output row at a time. Let $w \in \mathbb{R}^{C_{\text{in}}}$ denote one row of W_l, and let δw denote its perturbation after pruning. The importance-aware Hessian for this row is

$$H_{l,\text{imp}} = 2 \sum_{t=1}^{T} \alpha_t \mathbb{E}\left[X'_{l,t} X'^{\top}_{l,t} \right], \tag{10}$$

where the expectation is taken over the calibration data. The change in the layer-wise reconstruction objective is then

$$\Delta\mathcal{J} = \frac{1}{2} \delta w^{\top} H_{l,\text{imp}} \delta w. \tag{11}$$

To compute the saliency of weight w_q, we set this weight to zero while allowing the remaining weights in the row to compensate. The pruning constraint is

$$\mathbf{e}_q^{\top} \delta w + w_q = 0, \tag{12}$$

where $\mathbf{e}_q$ is the unit vector selecting coordinate q. We minimize $\Delta\mathcal{J}$ subject to this constraint using the Lagrangian

$$\mathcal{G}(\delta w, \lambda) = \frac{1}{2} \delta w^{\top} H_{l,\text{imp}} \delta w + \lambda \left(\mathbf{e}_q^{\top} \delta w + w_q \right). \tag{13}$$

Setting the first-order conditions to zero gives

$$\frac{\partial \mathcal{G}}{\partial \delta w} = 0 \implies H_{l,\text{imp}} \delta w + \lambda \mathbf{e}_q = 0, \tag{14}$$

$$\frac{\partial \mathcal{G}}{\partial \lambda} = 0 \implies \mathbf{e}_q^{\top} \delta w + w_q = 0. \tag{15}$$

Thus,

$$\delta w = -\lambda H_{l,\text{imp}}^{-1} \mathbf{e}_q. \tag{16}$$

Substituting this into the constraint yields

$$-\lambda \mathbf{e}_q^{\top} H_{l,\text{imp}}^{-1} \mathbf{e}_q + w_q = 0, \tag{17}$$

and therefore

$$\lambda = \frac{w_q}{\left[H_{l,\text{imp}}^{-1} \right]_{qq}}. \tag{18}$$

The optimal compensating update is

$$\delta w = -\frac{w_q}{\left[H_{l,\text{imp}}^{-1}\right]_{qq}} H_{l,\text{imp}}^{-1} \mathbf{e}_q. \tag{19}$$

The resulting saliency score is

$$S_{l,q} = \Delta\mathcal{J} = \frac{1}{2}\delta w^\top H_{l,\text{imp}} \delta w = \frac{w_q^2}{2\left[H_{l,\text{imp}}^{-1}\right]_{qq}}. \tag{20}$$

This derivation has the same structure as classical OBS (93). The difference is that the Hessian is computed from importance-filtered activations $X'_{l,t}$ rather than raw activations $X_{l,t}$. Thus, the saliency estimate remains an OBS-style reconstruction criterion, but it is biased toward preserving weights that affect spatially important regions.

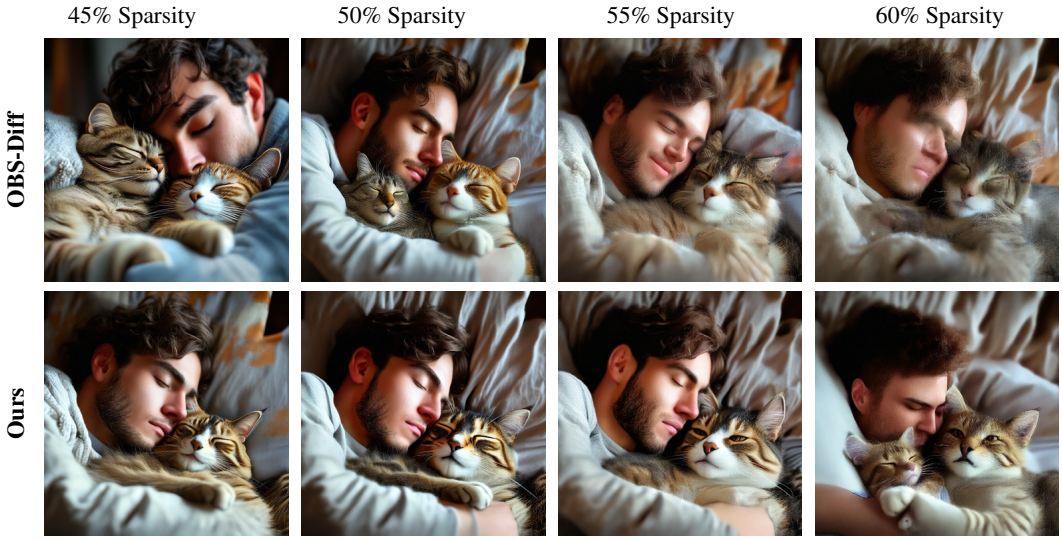


Figure 8: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A young man and his cute cat enjoy a nap together."



Figure 9: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A kitten sitting in a sink with a green brush with green bristles."



Figure 10: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A black Honda motorcycle parked in front of a garage."

D Broader Impact

This work contributes to making text-to-image diffusion models more efficient, practical, and accessible. By reducing model parameters while preserving perceptually important content, importance-aware pruning can lower memory usage, inference cost, and energy consumption. This can broaden access to high-quality generative models beyond large-scale compute environments, enabling more researchers, developers, and practitioners to use diffusion models on limited hardware.

A key positive impact of our work is that it provides a more flexible view of model compression. Instead of treating all generation errors equally, our framework allows pruning to prioritize the visual properties that matter for a given application, such as foreground objects, structural boundaries, or fine image details. This makes compression more adaptable to real deployment needs, where preserving



Figure 11: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A motorcycle is parked on a dirt road in a forest."



Figure 12: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A customized motorcycle with a large rear and skinny front tire."

semantically important regions is often more valuable than uniformly minimizing reconstruction error everywhere.

The proposed framework is also model-agnostic and easy to customize. Since the importance signal can come from readily available sources such as CFG responses, edge detectors, or object detectors, the method does not require retraining a separate evaluation model or redesigning the diffusion architecture. This makes it a practical tool for efficient deployment and opens a path toward application-aware compression, where different domains can define their own preservation priorities.

At the same time, importance-aware pruning introduces potential risks. If the importance signal is biased, incomplete, or poorly chosen, the compressed model may systematically preserve some visual concepts while degrading others. Category-targeted pruning can improve fidelity for selected categories, but it could also reduce performance on underrepresented or non-target categories if used



Figure 13: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "a shop a car traffic lights and buildings."



Figure 14: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A beautiful dessert waiting to be shared by two people."

carelessly. More generally, controllable compression creates a new axis of model specialization, which should be evaluated not only by average metrics but also by subgroup robustness, object-level fidelity, and failure cases.

This work does not introduce new generative capabilities or modify the training data of the underlying model. Instead, it focuses on improving the efficiency and fidelity of existing models under compression. As a result, its main expected benefits are reduced computational cost, improved deployment feasibility, and better preservation of user-relevant image content after pruning.

More broadly, our results suggest that efficient generative modeling should be evaluated not only by compression ratio, but also by what content is preserved after compression. We hope this encourages future work on controllable, interpretable, and task-aware compression methods for generative models, where efficiency gains are achieved without sacrificing the visual details most important to users.



Figure 15: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A cat eating a bird it has caught."

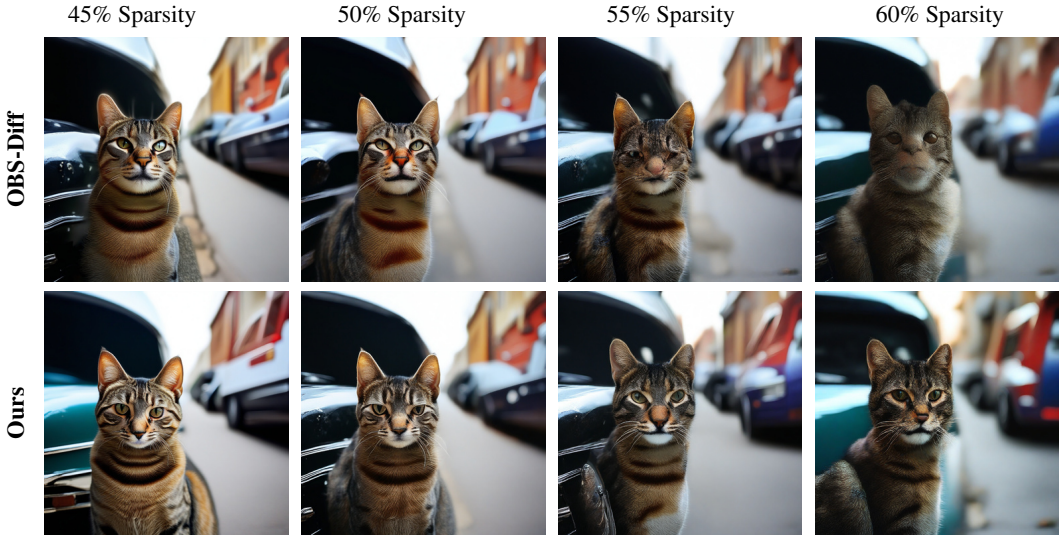


Figure 16: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A cat at attention between two parked cars."

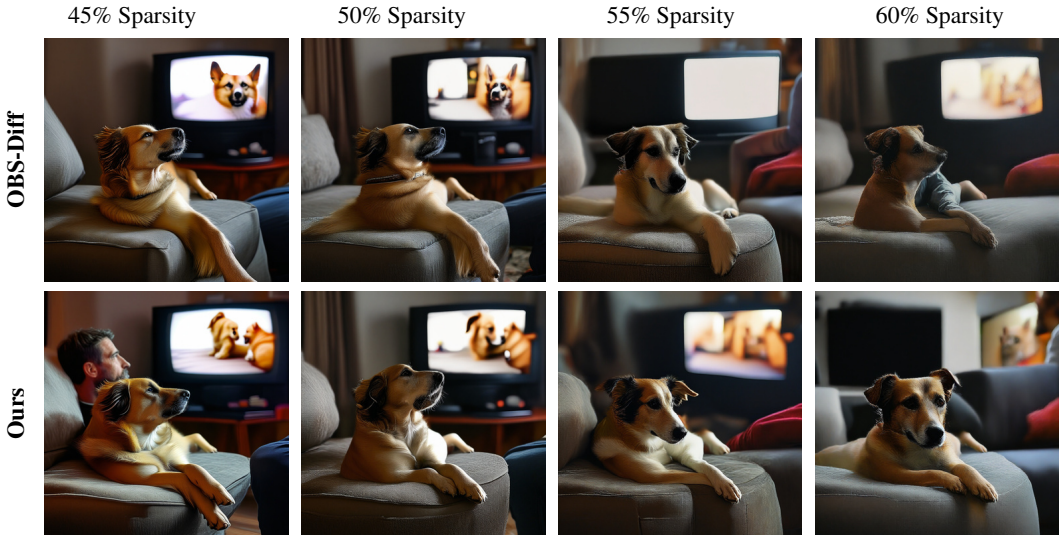


Figure 17: **Qualitative results on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A dog sitting between its masters feet on a footstool watching tv."



Figure 18: **Qualitative resutls on SD3-Medium.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 30%, 40%, 45%, and 50% unstructured sparsity. **Prompt:** "A man walking beside sheep on a country road."



Figure 19: **Qualitative resutls on SD3-Medium.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 30%, 40%, 45%, and 50% unstructured sparsity. **Prompt:** "An American Airlines plane is in the sky."



Figure 20: **Qualitative resutls on SD3-Medium.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 30%, 40%, 45%, and 50% unstructured sparsity. **Prompt:** "This new fridge goes great in this clean kitchen."



Figure 21: **Qualitative resutls on SD3-Medium.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 30%, 40%, 45%, and 50% unstructured sparsity. **Prompt:** "Two people riding a motorcycle to the beach."

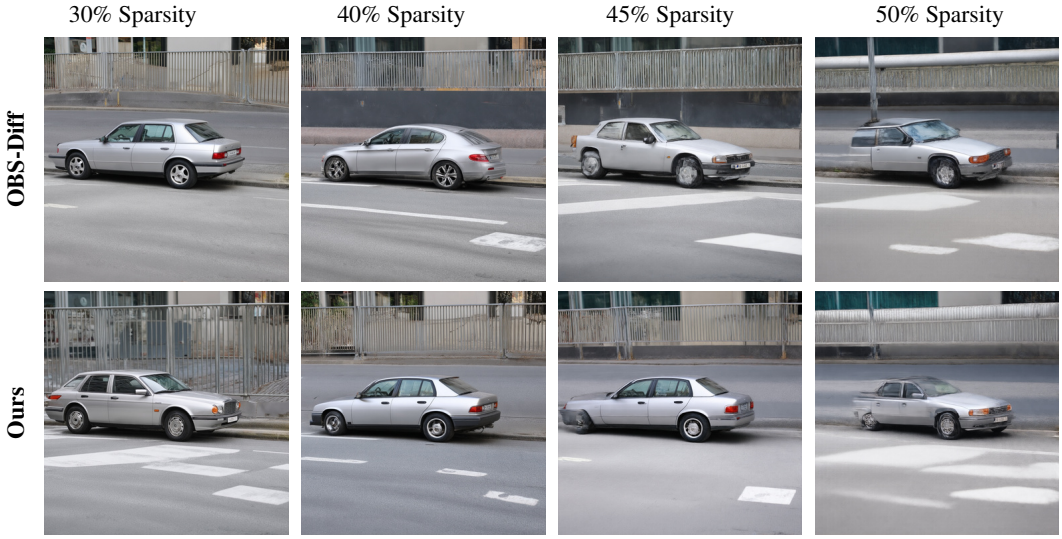


Figure 22: **Qualitative resutls on SD3-Medium.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 30%, 40%, 45%, and 50% unstructured sparsity. **Prompt:** "A silver car in the street next to a metal railing."



Figure 23: **Target-Category Pruning: woman.** Comparison of general and targeted pruning baselines (OBS-Diff) against our importance-aware variants at 60% sparsity level.

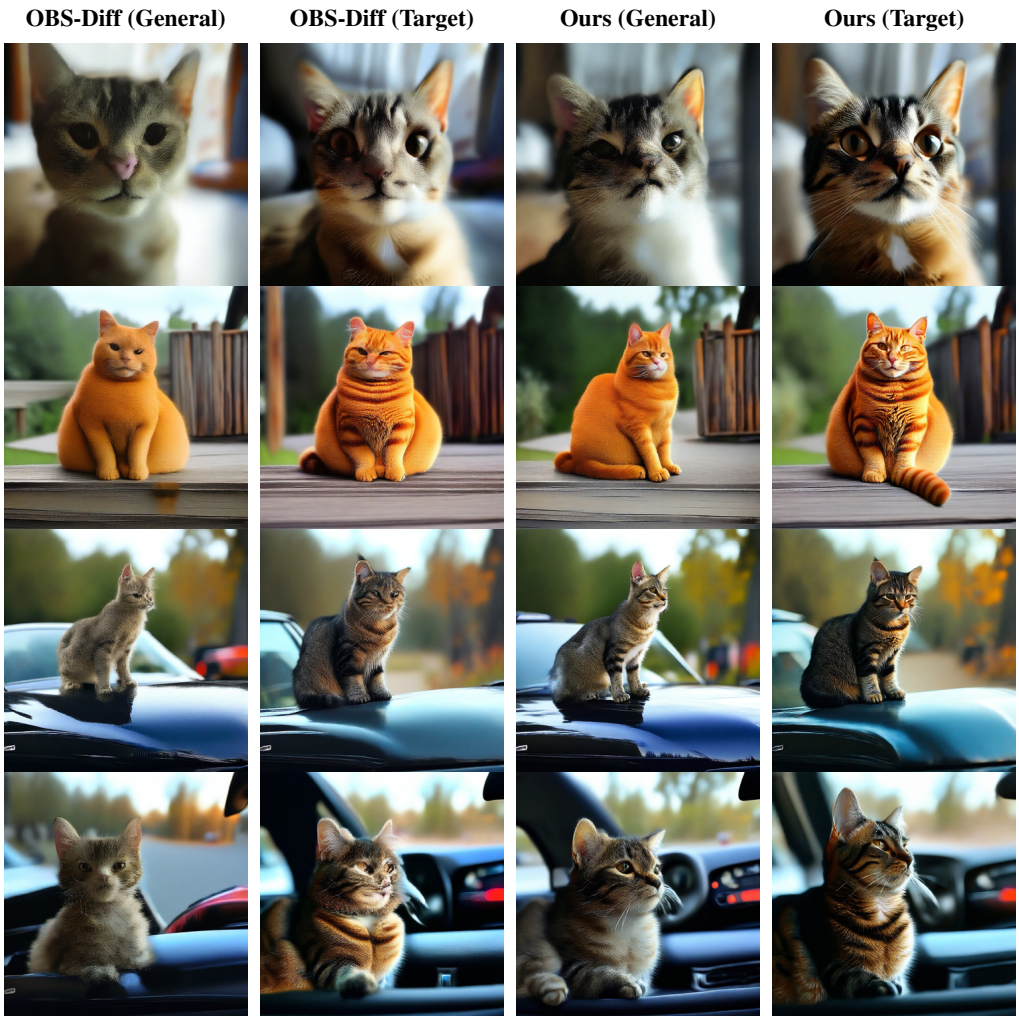


Figure 24: **Target-Category Pruning: cat.** Comparison of general and targeted pruning baselines (OBS-Diff) against our importance-aware variants at 60% sparsity level.

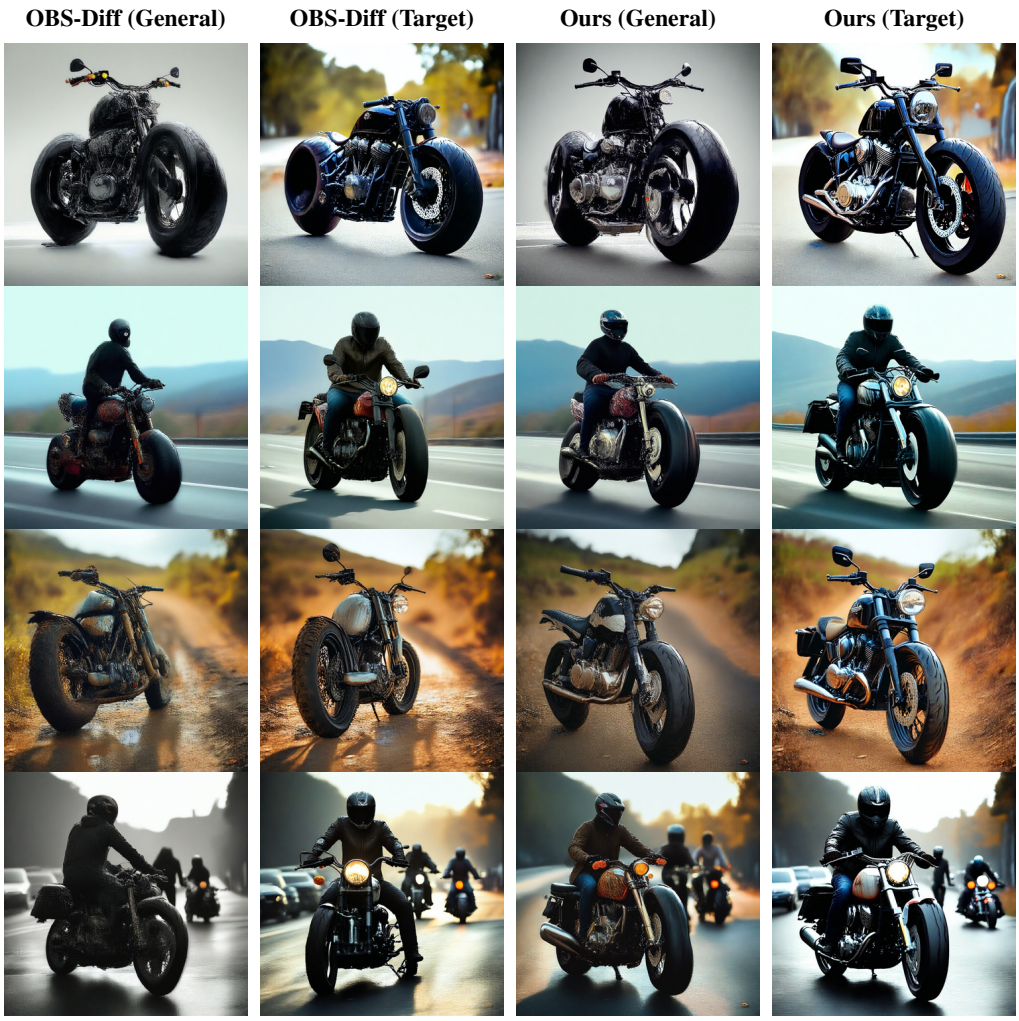


Figure 25: **Target-Category Pruning: motorcycle.** Comparison of general and targeted pruning baselines (OBS-Diff) against our importance-aware variants at 60% sparsity level.



Figure 26: **Target-Category Pruning: airplane.** Comparison of general and targeted pruning baselines (OBS-Diff) against our importance-aware variants at 60% sparsity level.



Figure 27: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A black Honda motorcycle parked in front of a garage."



Figure 28: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A Honda motorcycle parked in a grass driveway."



Figure 29: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A shot of an elderly man inside a kitchen."

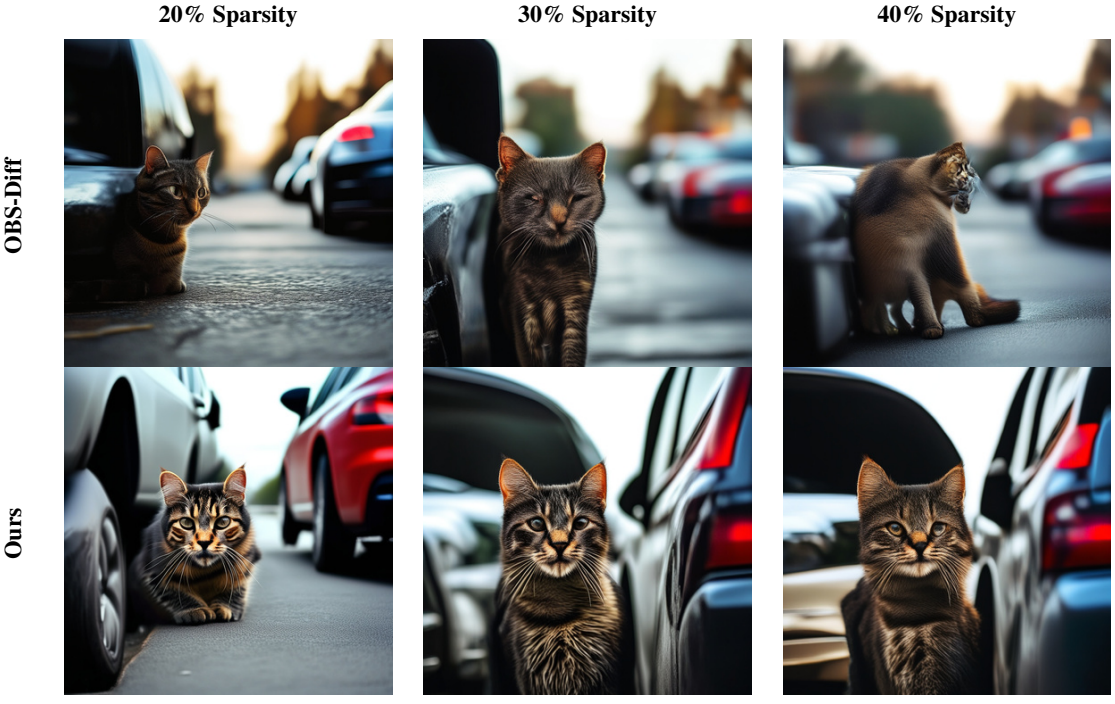


Figure 30: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A cat in between two cars in a parking lot."

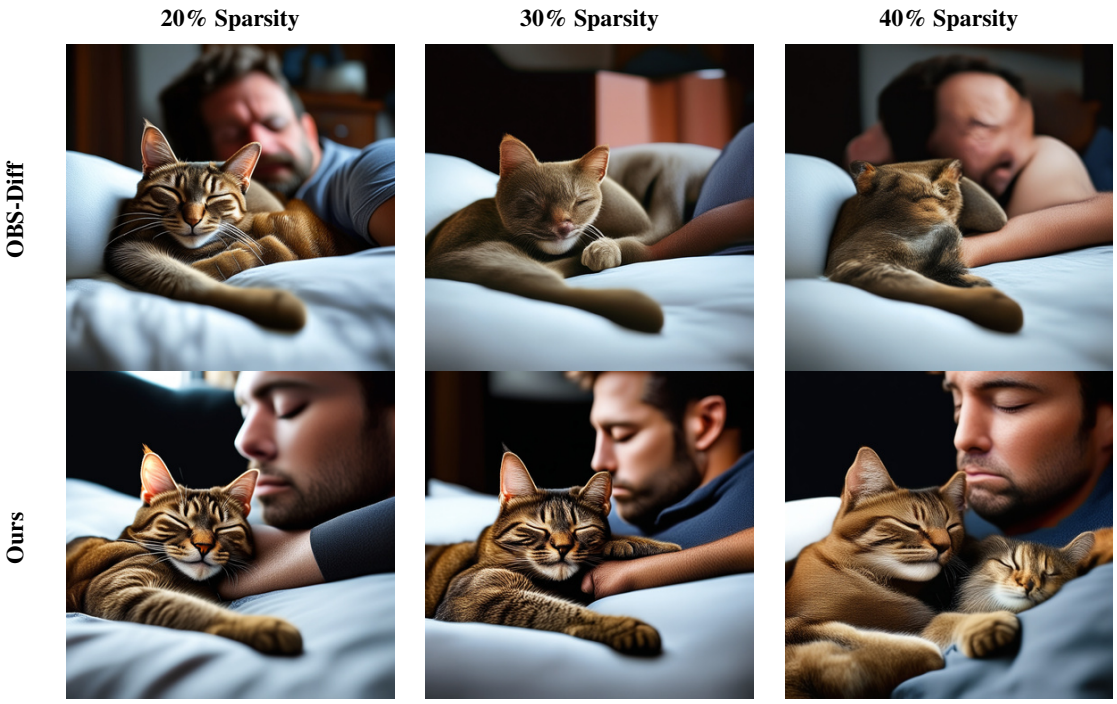


Figure 31: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "a man sleeping with his cat next to him"



Figure 32: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A cat stands between two parked cars on a grassy sidewalk."

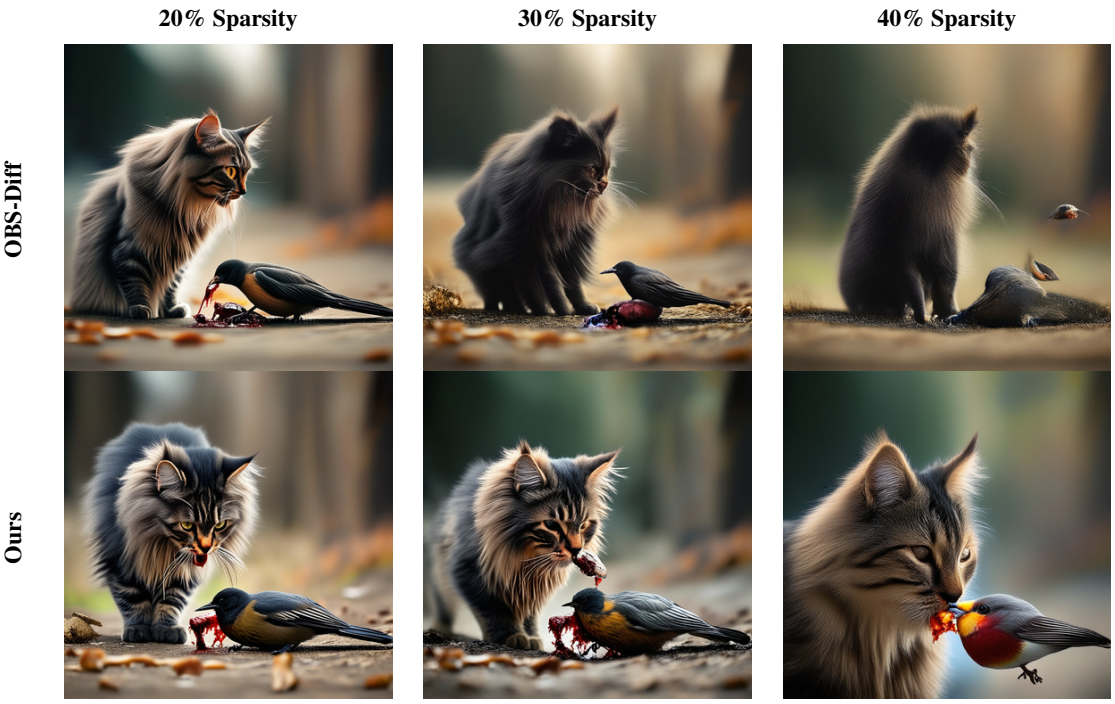


Figure 33: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A long haired cat eating a dead bird."

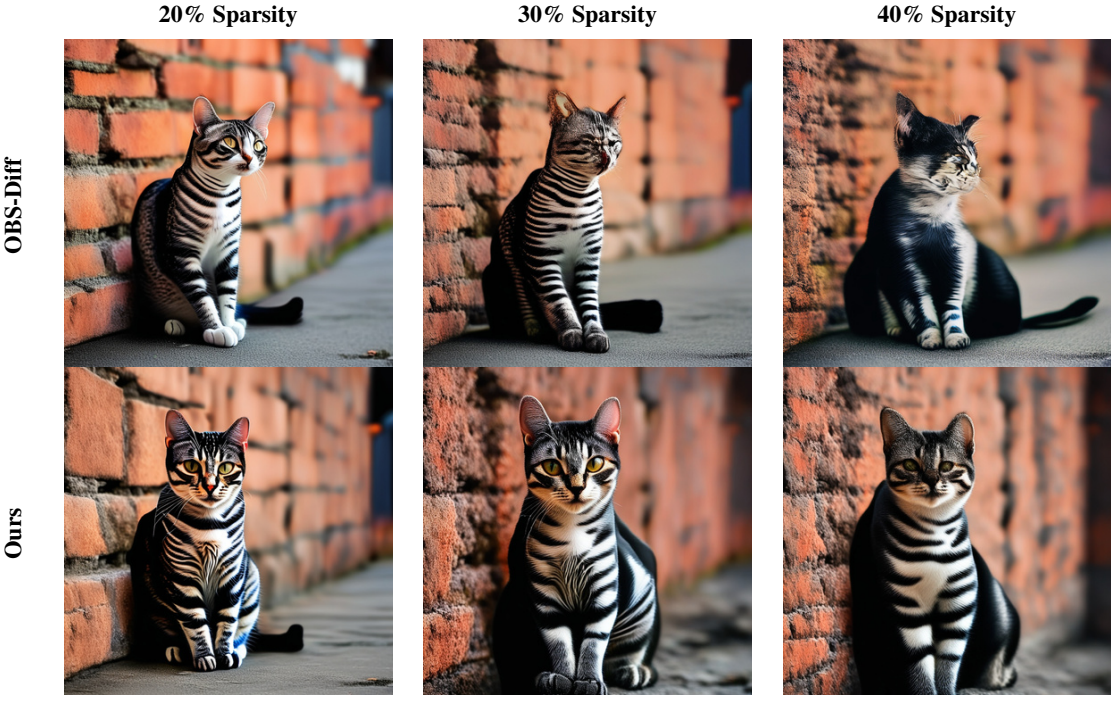


Figure 34: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "a stripped cat sitting near a brick wall"

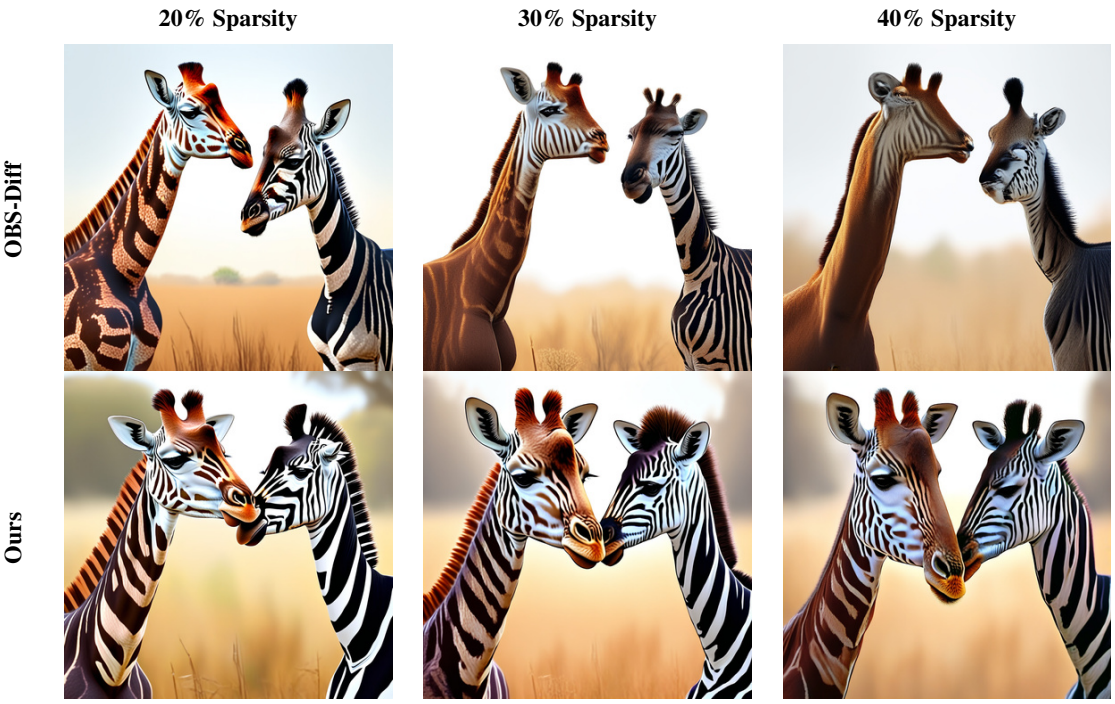


Figure 35: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A giraffe and a zebra checking each other out."



Figure 36: **Qualitative resutls on PixArt-Σ.** Comparison between OBS-Diff and our importance-aware pruning guided by CFG (2) signals at 20%, 30%, and 40% structured sparsity. **Prompt:** "A red truck has a black dog in the drivers chair."

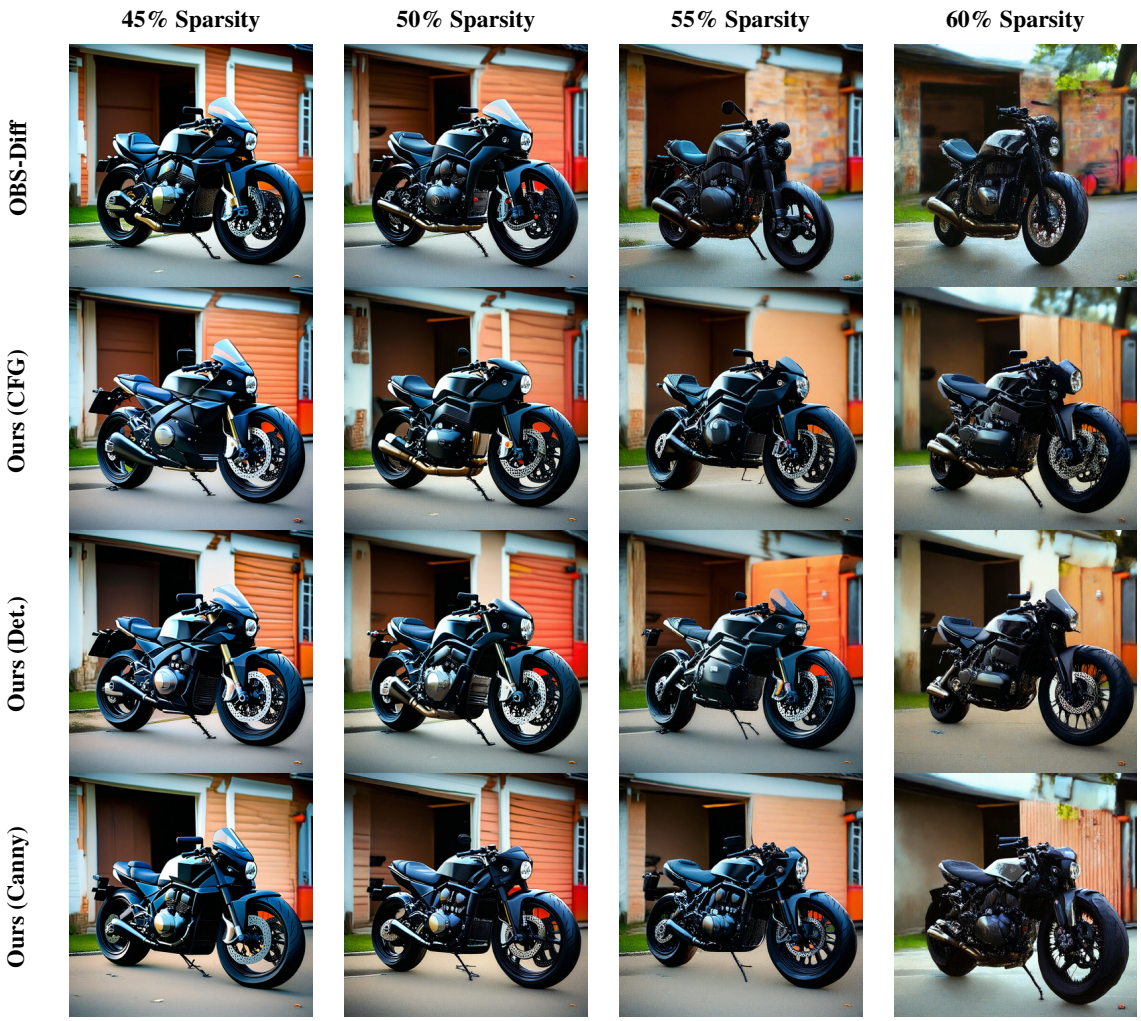


Figure 37: **Different importance signals.** Comparison between OBS-Diff and our importance-aware variants guided by CFG (2), CFG+ object-detection(14), and Canny-edge(13) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A black Honda motorcycle parked in front of a garage."



Figure 38: **Different importance signals.** Comparison between OBS-Diff and our importance-aware variants guided by CFG (2), CFG+ object-detection(14), and Canny-edge(13) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A large plant is in the corner of a small bathroom."

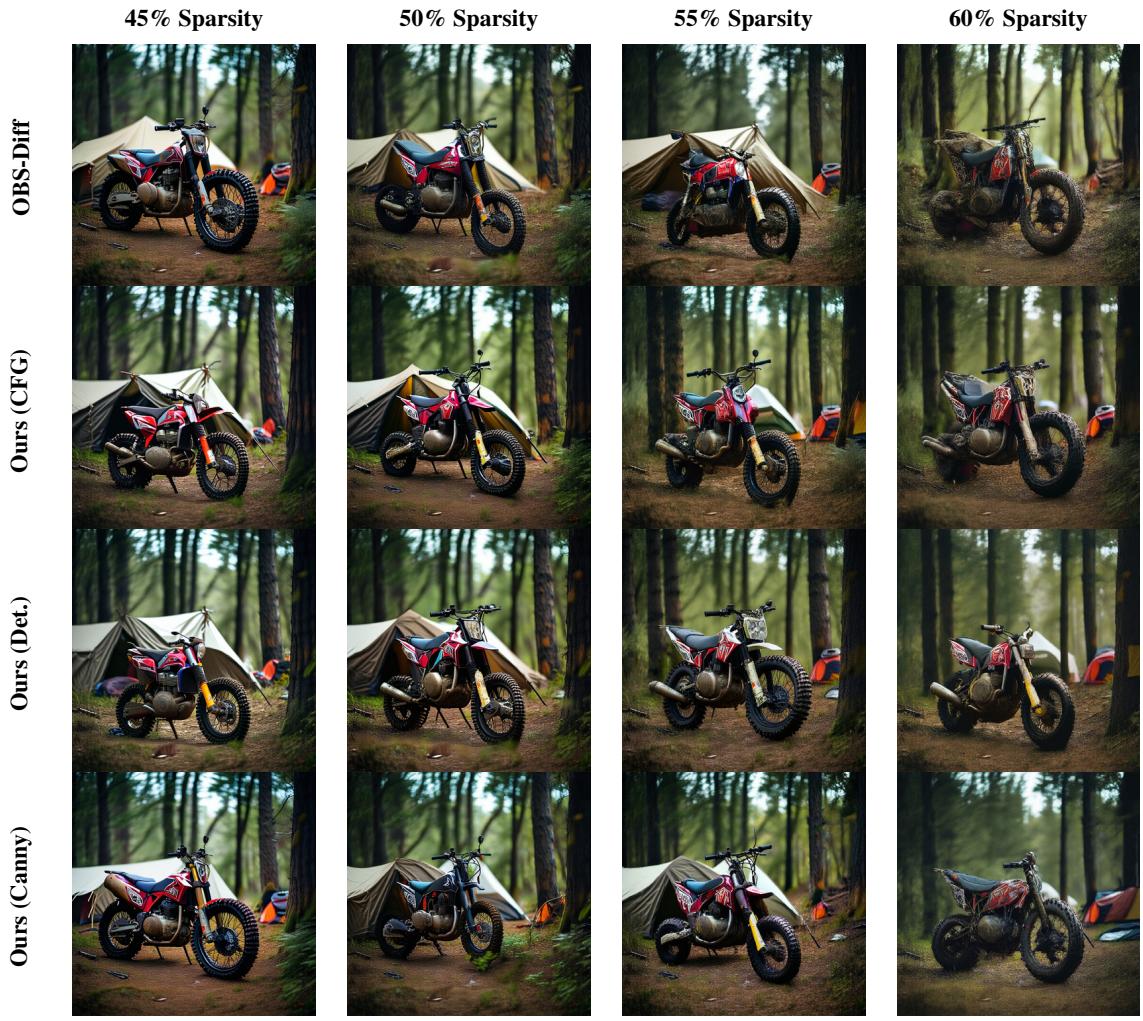


Figure 39: **Different importance signals.** Comparison between OBS-Diff and our importance-aware variants guided by CFG (2), CFG+ object-detection(14), and Canny-edge(13) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A dirt bike parked near a tent in the woods."

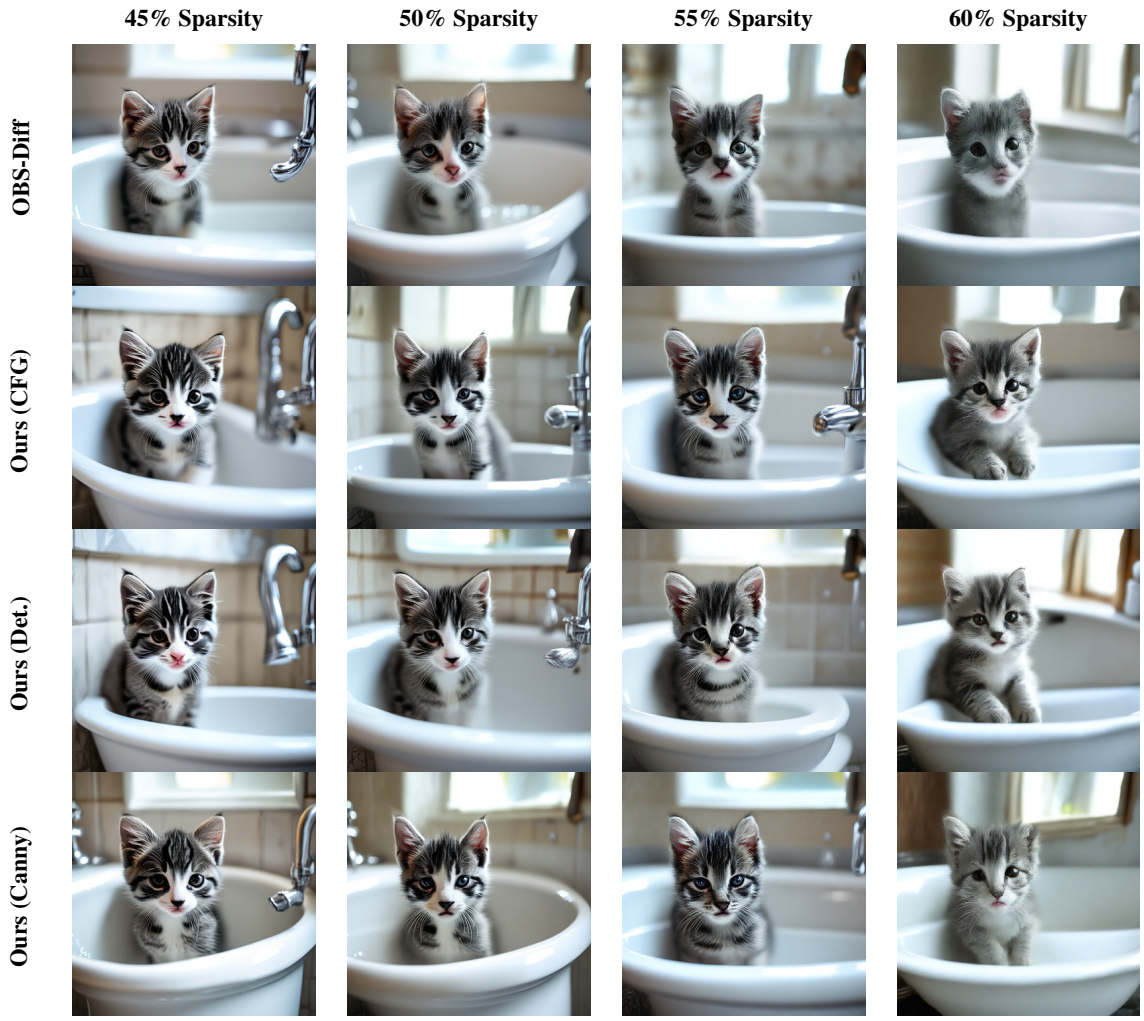


Figure 40: **Different importance signals.** Comparison between OBS-Diff and our importance-aware variants guided by CFG (2), CFG+ object-detection(14), and Canny-edge(13) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A gray and white kitten in a white bathroom sink."



Figure 41: **Different importance signals.** Comparison between OBS-Diff and our importance-aware variants guided by CFG (2), CFG+ object-detection(14), and Canny-edge(13) signals at 45%, 50%, 55%, and 60% unstructured sparsity. **Prompt:** "A large truck is parked beside an old, wooden bench."

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the contributions and scope of the paper. We state that the paper introduces an importance-aware pruning framework for diffusion models and support this claim through the proposed formulation, derivation, and empirical evaluation. The claims are limited to the evaluated diffusion models, pruning settings, and importance signals considered in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the proposed approach, including its dependence on the quality of the importance signal, the computational cost of estimating pruning statistics, and the finite scope of evaluation across diffusion models and sparsity regimes. We also note that performance may degrade at very high sparsity levels.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides the derivation of the importance-aware OBS saliency score, including the weighted reconstruction objective, the corresponding Hessian estimate, and the resulting closed-form pruning criterion. The derivation follows the standard local quadratic approximation used in OBS-style pruning and is included in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the evaluated diffusion backbones, pruning baselines, sparsity levels, calibration procedure, importance signals, evaluation prompts, and evaluation metrics. These details provide the information needed to reproduce the main experimental results that support the paper's claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The submission does not currently provide public code and scripts for reproducing all main experimental results. We describe the method and experimental setup in the paper and plan to release code and reproduction instructions upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the diffusion models, pruning settings, calibration data, sparsity levels, importance-guidance construction, baselines, and evaluation metrics. Since the method is post-training pruning, the relevant details are the pruning, calibration, saliency computation, and evaluation protocols, which are described in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports quantitative results over benchmark prompts and image-quality metrics, but does not consistently provide error bars, confidence intervals, or statistical significance tests for all main experimental comparisons.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not yet fully specify the compute resources required for each experiment, such as GPU type, GPU memory, runtime per pruning/evaluation run, and total compute budget. We will include these details in the final version.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. It uses existing diffusion models and evaluation data for model compression, and does not involve human-subject experiments, private data collection, deception, or unsafe data release.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts from reducing the computational and memory cost of diffusion models, which can make generative models more efficient and accessible. It also acknowledges possible negative impacts, since improving the efficiency of image generation may lower the cost of misuse, including synthetic media generation and disinformation.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new high-risk pretrained generative model or scraped dataset. The work focuses on a pruning method for existing diffusion models, and any use of existing models follows their original release conditions and safety restrictions.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original works for the diffusion backbones, pruning baselines, datasets or prompt sets, and evaluation metrics used in the experiments. We use existing assets according to their released terms and licenses. The final version will explicitly list model and dataset licenses where applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or pretrained model as a primary asset. The main contribution is the pruning method and its evaluation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See Appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: There is no potential risk that we are aware of.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use LLMs as an important, original, or non-standard research component. Any use of language tools for writing, editing, or formatting does not affect the scientific methodology, experiments, or originality of the work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.