Rethinking Visual Autoregressive Sampling with Information-Grounding Guidance

Ky Dan Nguyen [1 *] **Hoang Lam Tran** [1 *] **Anh-Dung Dinh** [1] **Daochang Liu** [2] **Weidong Cai** [1] **Xiuying Wang** [1] **Chang Xu** [1]

arXiv:2509.23876v3 [cs.CV] 4 Jul 2026

Abstract

Autoregressive (AR) models based on next-scale prediction have emerged as a powerful tool for image generation, but they face a critical weakness: information inconsistencies between patches across timesteps introduced by progressive resolution scaling. These inconsistencies scatter guidance signals, causing them to drift away from salient regions within the image and leaving behind ambiguous, unfaithful features during sampling. We tackle this challenge with Information-Grounding Guidance (IGG), a novel framework that anchors guidance to semantically important tokens via an attention-based dynamic weighting formulation, consequently ensuring that guidance and semantic contents remain tightly aligned. Across both class-conditioned and text-to-image generation tasks, IGG delivers sharper, more coherent, and semantically grounded images, demonstrating its efficacy for correcting AR-based methods. Our code is available at https://github.com/dnngky/info ground-guidance.

1. Introduction

Autoregressive (AR) modelling (Chen et al., 2020; Esser et al., 2021; Ramesh et al., 2021; Li et al., 2024b; Tian et al., 2024; Zhang et al., 2024; Tang et al., 2025; Han et al., 2025; Voronov et al., 2025) has established itself within the field of image generation due to its ability to produce high-quality outputs. AR models, at their core, sample image patches from joint distributions of discrete tokens, allowing generation of complex visual patterns.

A notable advancement in AR modelling is the introduction of multi-scale tokenisation strategies (Tian et al., 2024; Han et al., 2025; Voronov et al., 2025; Zhang et al., 2024; Tang et al., 2025), which enable the models to capture features from coarse to fine levels. This hierarchical approach better aligns with the inherent multi-scale structure of natural images, allowing AR models to effectively model both global structures and fine-grained details. Building on this paradigm, Tian et al. (2024) redefined the autoregressive process as "next-scale" prediction, diverging from "next-token" prediction, which has long been the standard of AR modelling. This methodology—henceforth referred to as *scale-wise* AR (SwAR) modelling as per Voronov et al. (2025); Ren et al. (2024)—is able to generalise features across varying resolutions while effectively reducing the computation of the sampling step, offering a competitive alternative to diffusion-based approaches in terms of sampling quality and time.

Regardless of modelling approaches, the sampling process of generative models relies heavily on various guidance techniques (Kynkäänniemi et al., 2024; Hong et al., 2023; Ho & Salimans, 2021; Dinh et al., 2023a;b; 2025; Karras et al., 2024a). In diffusion-based models, guidance is widely utilised to improve image fidelity at the expense of sample diversity, typically by modulating the denoising trajectory with additional conditioning signals. This mechanism has proven effective in improving visual quality and semantic alignment, particularly in class-conditioned generation tasks. Inspired by the efficacy of diffusion-based guidance, some techniques have been adopted for SwAR modelling. Most notably, Tian et al. (2024) employed classifier-free guidance (Ho & Salimans, 2021) on token predictions (instead of noise and score predictions in diffusion models). Nevertheless, despite impressive empirical results, AR-based guidance is considerably less understood compared to its diffusion-based counterpart. This gap is especially prominent given the fundamental differences between diffusion and AR models.

In light of this, this work is dedicated to investigating the behaviour of AR-based guidance, with a focus on SwAR modelling. In particular, our analysis reveals that guidance in SwAR models is often misaligned, in stark contrast to

*Equal contribution [1]The University of Sydney, Sydney, New South Wales, Australia [2]The University of Western Australia, Perth, Western Australia, Australia. Correspondence to: Chang Xu <c.xu@sydney.edu.au>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

guidance in diffusion models. From this key insight, we propose IGG (**I**nformation-**G**rounding **G**uidance), a novel guidance scheme designed to accentuate this behaviour, which we posit improves the overall sampling quality of SwAR models. In particular, IGG infers the semantical importance of each token from its surrounding context and applies guidance accordingly. To the best of our knowledge, it is the first method specifically designed for SwAR modelling. We evaluate IGG on various generative tasks. The results of these experiments consistently demonstrate improvements over baselines, thereby validating both our hypothesis and the efficacy of our method. Our contributions in this paper are:

1. Present a technical analysis on the dynamics of guidance in diffusion and SwAR models, revealing that guidance tends to be semantically misaligned in the latter.
2. Propose IGG, a novel technique that adaptively concentrates guidance on semantically important tokens.
3. Demonstrate the effectiveness of IGG through extensive experiments on class-conditioned generation and text-to-image generation.

2. Related Work

Diffusion models. Concurrent to the development of AR-based generative models, denoising diffusion models have been receiving much attention for its impressive image generation capability (Rombach et al., 2022; Peebles & Xie, 2023; Karras et al., 2024b). At its core, diffusion-based models learn how to make progressive denoising steps that transform pure noise into the target image. To enhance the quality of the generated samples at the trade-off of diversity, diffusion models rely on guidance methods (see below).

Autoregressive (AR) models. AR models have been highly regarded for their success in language tasks. Prior adaptations of these transformer-based architectures for generative tasks Chen et al. (2020); Esser et al. (2021) have been exploring the use of transformers to generate image patches in a raster-scan order. Masked autoregressive models Chang et al. (2022); Li et al. (2024b); Fan et al. (2025) changed this generation behaviour by sampling image patches in a random order at each step. More notably, autoregressive models using the "next-scale prediction" paradigm pioneered by Tian et al. (2024) have been gaining attention for their competitive generation quality to diffusion models.

Classifier-free guidance (CFG). Amongst the breadth of guidance methods for generative modelling (Dhariwal & Nichol, 2021; Ho & Salimans, 2021; Hong et al., 2023; Dinh et al., 2023a;b; 2025; Kynkäänniemi et al., 2024; Karras et al., 2024a), CFG (Ho & Salimans, 2021) has remained the de facto technique which has inspired many modern variants in and beyond computer vision, most notably NLP Li et al. (2025); Liao et al. (2025); Zhang et al. (2025). At its core, CFG uses an unconditional model as the distribution to steer the class-conditioned version away from. This mechanism is inspired by classifier guidance (Dhariwal & Nichol, 2021) which opts for an explicit classifier to guide the denoising process. Recent advancements have seen the application of CFG to discrete diffusion and flow models (Nisonoff et al., 2025; Schiff et al., 2025). Given its demonstrated generalisability, CFG has also been adapted to SwAR modelling (Tian et al., 2024) where its efficacy has been empirically verified.

3. Background

Scale-wise autoregressive modelling. Given a vocabulary $\mathcal{V}$, a *token map* s_k is defined as a set of $h_k w_k$ tokens $\{t_{1,1}, t_{1,2}, \cdots, t_{h_k,w_k}\} \subseteq \mathcal{V}$ or, more intuitively, a $h_k \times w_k$ grid of tokens. In its general form, a SwAR model constructs a series of token maps $s := (s_1, \cdots, s_K)$ using an encoder, such that $h_K \times w_K$ matches the resolution of x. Then, a model is trained to predict s in an autoregressive manner, maximising the likelihood $p(s|c) = \prod_{k=1}^{K} p(s_k|s_{<k}, c)$, where c denotes the conditioning signal. Finally, a reconstruction of the original x is produced from the predicted token maps using a decoder.

SwAR-based classifier-free guidance. The mechanism of classifier-free guidance in SwAR modelling is analogous to diffusion-based modelling, with the only exception in what is being guided (token predictions, in contrast to noise/score predictions). In particular, guidance is applied on each inference step by interpolating between the outputs of the unconditioned model $p_\theta(\cdot)$ and its conditioned counterpart $p_\theta(\cdot|c)$, which may be expressed as

$$\tilde{p}_\theta(s_k|c) = (1+\gamma_k)\, p_\theta(s_k|c) - \gamma_k\, p_\theta(s_k), \qquad (1)$$

where $\gamma \in \mathbb{R}$ is the guidance scale. Following guidance, sampling can be performed on $p_\theta(s_k|c)$ to obtain the final prediction $\hat{s}_k \in \mathcal{V}^{h_k \times w_k}$. Interestingly, while the original diffusion-based implementation (Ho & Salimans, 2021) fixes the guidance scales $\gamma_1 = \cdots = \gamma_K$, it is possible to generalise it to a guidance *schedule*. For example, Tian et al. (2024) defined the schedule $\gamma_k := w \cdot k/(K-1)$, where $w \in \mathbb{R}$ is a hyperparameter, to gradually accentuate guidance throughout inference.

4. Analysing the Behaviour of CFG

This section aims to elucidate the key behaviours of CFG and differences between the guidance dynamics of diffu-

sion modelling as opposed to AR modelling, thereby shedding some light on the gap in performance between these two approaches in practice. We begin by offering an alternative perspective to the conventional interpretation of CFG as interpolating between unconditioned and conditioned predictions:

$$\tilde{p}_\theta(s_k|c) = p_\theta(s_k) + \gamma_k \underbrace{\left(p_\theta(s_k|c) - p_\theta(s_k)\right)}_{p_\theta^{\rightarrow}(s_k|c)}. \quad (2)$$

Equation 2 becomes equivalent to Equation 1 by setting $\gamma_k := 1 + \gamma_k$. Under this formulation, CFG can be seen as "nudging" the unconditioned predictions towards the conditioned predictions. Indeed, this is consistent with the remark of the authors of the original implementation of CFG (Ho & Salimans, 2021). For ease of notations, we henceforth use $p_\theta^{\rightarrow}(\cdot|c)$ to denote the guidance signals characterising these "nudges". A visualisation of these guidance signals is depicted in Figure 2.

Guidance does not treat tokens equally. To validate our first hypothesis, we used *Pielou's evenness index* (Pielou, 1966)—or the normalised Shannon entropy—to evaluate the evenness of the guidance distribution associated with each token map,

$$\mathrm{PEI}(s_k|c) = \frac{H(p_\theta^{\rightarrow}(s_k|c))}{\ln(h_k w_k)}, \quad (3)$$

where H denotes Shannon entropy. Note that $\mathrm{PEI}(\cdot) \in [0, 1]$, where a greater value indicates a more even distribution. In our evaluation, for each image, we compute $\mathrm{PEI}(s_k|c)$ for every sampling step k (although $k = 1$ is omitted for SwAR models because it makes little sense to evaluate a distribution comprising a single value). The overall PEI score for an image is computed as the weighted mean where each sampling step is weighted by its corresponding output resolution (see Figure 1 for a visual illustration). For diffusion models, since the output resolution is constant throughout the sampling process, the weighted mean reduces to the unweighted mean. Comprehensive comparison between a representative model from each paradigm, reported in Table 1, reveals that diffusion modelling exhibits highly uneven guidance signals while SwAR modelling remains lacklustre.

Guidance prioritises semantically important tokens. Let $s_{fg} \subset s$ denote the set of foreground tokens representing semantically important regions of the generated image at an arbitrary scale. To evaluate whether guidance signals preferentially target these regions, we contrast the guided distribution $p := p_\theta^{\rightarrow}(s|c)$ with the synthetic unguided distribution $q := p_\theta^{\rightarrow}(s^*|c)$, where s^* is sampled from the background set $s_{bg} = s \setminus s_{fg}$. We quantify the divergence between p and q using the Jensen-Shannon distance:

$$\mathrm{JSD}(p, q) = \sqrt{\tfrac{1}{2}\mathrm{KL}(p \parallel m) + \tfrac{1}{2}\mathrm{KL}(q \parallel m)}, \quad (4)$$

Table 1. Comparison of the evenness (Evn, Equation 3) and divergence (Div, Equation 4) of guidance in EDM2 (Karras et al., 2024b) and VAR (Tian et al., 2024). Results are averaged scores obtained through ImageNet 512×512 class-conditioned generation. **Our method closes the gap between SwAR and diffusion models.**

Model	Evn (↓)	Div (↑)
EDM2-S	0.486	0.983
EDM2-XXL	0.560	0.964
VAR-$d36$-CFG	0.741	0.623
VAR-$d36$-IGG	0.665	0.751

where $m = \frac{1}{2}(p + q)$ represents the mixture distribution. Higher values of $\mathrm{JSD}(p, q) \in [0, 1]$ indicate a more pronounced semantic bias within the guidance signals. For our experiments, s_{fg} is identified via automatic segmentation (e.g., using YOLOv11 (Khanam & Hussain, 2024)), with the full procedure detailed in Appendix A. Results in Table 1 demonstrate significant divergence in diffusion models, whereas the SwAR model exhibits minimal separation, suggesting a lack of semantic prioritization.

Guidance is misaligned in SwAR models. In addition to the above quantitative evaluations, we also qualitatively compared the guidance dynamics of the two generative modelling paradigms. Visualisation of sampled images from SwAR model revealed numerous instances where guidance signals became progressively dispersed at increasing scale levels, leading to adversarial features in generated images. On the other hand, the diffusion model exhibits guidance signals that are consistently sharp and aligned to semantically important tokens (Figure 1).

5. Information-Grounding Guidance

Building on the observations in Section 4, we hypothesise that the potential of CFG has not been fully utilised in SwAR models. In SwAR models, guidance signals often weakens progressively with each sampling step, making it disadvantageous to directly apply guidance strategies deployed in diffusion models, as illustrated in Figure 1. This gap can be narrowed by tuning the guidance signals of AR models so that they more closely align with the evenness and divergence patterns observed in diffusion models. Such an effect could be achieved through alternative guidance methods that selectively emphasise certain important regions or foreground features of the image.

In light of this, we introduce a new guidance framework based on the key observation from Section 4 that **not all tokens should be equally guided**. As alluded to in the previous paragraph, our method needs to identify semantically important features at inference time and amplify their corresponding guidance signals. To this end, our framework is

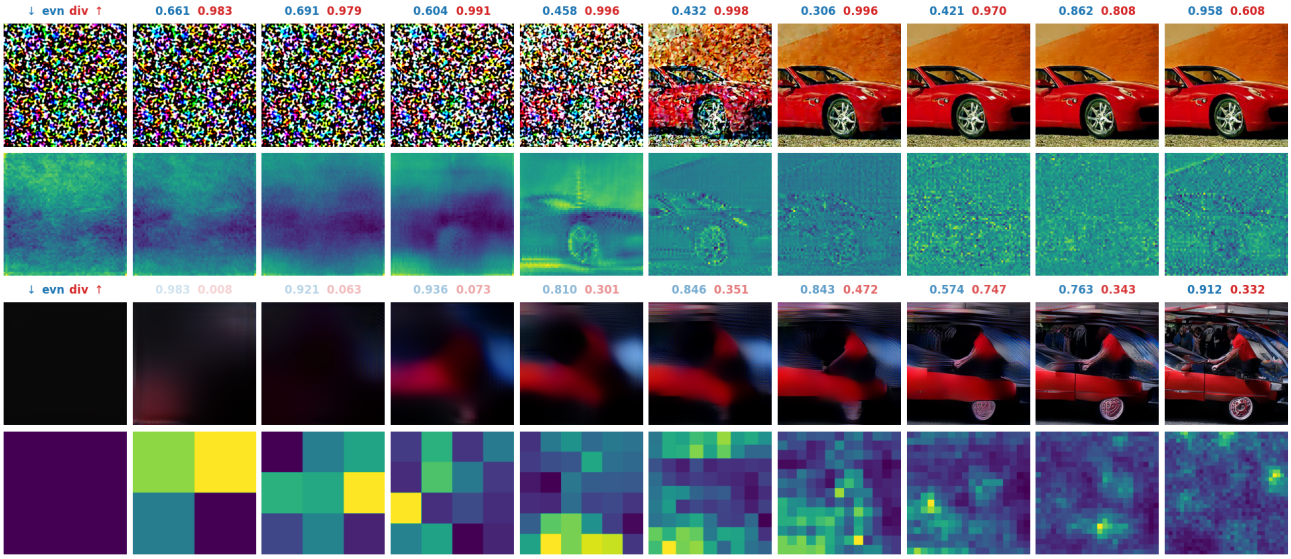


Figure 1. Distribution of guidance throughout the sampling process of EDM2 (Karras et al., 2024b) (top) and VAR (Tian et al., 2024) (bottom) on ImageNet 512 × 512 class-conditioned generation (class: sports car). For the sake of comparison, the original number of sampling steps of EDM2 (32 steps) has been modified to match VAR. Sampling steps are respectively labelled with their **evenness** and **divergence** scores, with opaqueness corresponding to relative contributions to the weighted mean score. **While EDM2 exhibits guidance signals that are sharp and consistently aligned to foreground objects, VAR exhibits guidance signals that are poorly aligned and becomes progressively fainter.** Additional examples can be found in Appendix B, and a visualisation of guidance in EDM2 across the full 32 sampling steps in Appendix D.

defined as

$$\tilde{p}_\theta(s_k|c) = p_\theta(s_k) + f_k(s_k|c) \cdot p_\theta^{\rightarrow}(s_k|c), \tag{5}$$

where $f_k : \mathcal{V}^{h_k \times w_k} \rightarrow \mathbb{R}^{h_k \times w_k}$ is a function that assigns a (possibly unique) weight to each token in s_k (according to their relative importance). Accordingly, this framework subsumes CFG as a special case where $f_k(s_k|c) = \gamma_k$.

To realise f_k, we observe that tokens representing the same region tend to carry similar values, and a token surrounded by other salient tokens should itself be considered important. This naturally leads us to the attention mechanism, which is well-suited for capturing such contextual relationships. Prior work has demonstrated that attention is highly effective for selecting informative tokens across diverse vision tasks—including classification Ryoo et al. (2021), captioning You et al. (2016), and in-painting Liu et al. (2019). In particular, Ryoo et al. (2021) showed that pairwise attention can automatically highlight critical visual tokens while maintaining efficiency comparable to state-of-the-art methods. In our framework, however, it is more appropriate to apply attention not to the tokens themselves but to the guidance signals $p_\theta^{\rightarrow}(\cdot|c)$, since these directly steer the sampling process. Accordingly, we implement f_k as a self-attention operation over these guidance signals:

$$f_k(s_k|c) = \gamma_k \cdot \text{softmax}\left(\frac{p_\theta^{\rightarrow}(s_k|c)[p_\theta^{\rightarrow}(s_k|c)]^T}{\sqrt{|\mathcal{V}|}}\right). \tag{6}$$

Note that the guidance scale γ_k is retained to facilitate guidance tuning on a global level. Plugging Equation 6 into Equation 5 yields our proposed guidance scheme, which we term *Information-Grounding Guidance* (IGG).

6. Experimental Results

This section presents extensive evaluations of IGG and quantitative analysis on the metrics proposed in Section 4. For class-conditioned image generation, we sampled 50,000 images across 1,000 classes provided by ImageNet (Deng et al., 2009). For text-to-image generation, we conducted experiments on three prompt sets—MJHQ, MS-COCO, and GenEval—sampling 1 image per prompt for the first two set and 4 images per prompt for the last set.

6.1. Class-Conditioned Image Generation

Quantitative results. To assess the scalability of IGG, we apply it to guide the generation of both 256 × 256 and 512 × 512 images, with VAR Tian et al. (2024) as the backbone model. To further evaluate its off-the-shelf practicality, we directly used pre-trained models provided by the authors on HuggingFace[1]. Our results in Table 2 demonstrate that IGG consistently outperforms CFG on VAR and numerous diffusion models in terms of both FID and IS. Furthermore, we observe that the superiority of IGG is more pronounced for 512 × 512 image generation. We attribute this phenomenon to the notion that VAR-$d36$ is required to predict larger token maps and, consequently, leaves more

[1] https://huggingface.co/FoundationVision/var.

Table 2. Comparison of IGG against guidance techniques on ImageNet class-conditioned generation task using representative diffusion (Dhariwal & Nichol, 2021; Rombach et al., 2022; Peebles & Xie, 2023) and SwAR models (Tian et al., 2024). Included guidance methods include *classifier guidance* (CLSG, Dhariwal & Nichol (2021)), *entropy-driven sampling* (EDS, Zheng et al. (2022)), *PixelAsParam* (PXP, Dinh et al. (2023a)), *classifier-free guidance* (CFG, Ho & Salimans (2021)), *progressive guidance* (PROG, Dinh et al. (2023b)), and *representative guidance* (REPG, Dinh et al. (2025)). Wherever possible, we report the guidance schedule scale w (see Section 3). Note that all models but VAR used a fixed guidance schedule. ↑/↓ indicates that higher/lower is better. Best results for each resolution are **bolded**. **Our method enhances the performance of VAR.**

Resolution	Model	Guidance	FID (↓)	IS (↑)	Pre (↑)	Rec (↑)
256×256	ADM	CLSG (w=1.00)	4.59	186.70	0.82	0.52
		EDS	4.09	221.57	0.83	0.50
		PXP	4.00	216.11	0.81	0.53
		CFG	3.76	191.31	0.77	0.53
		PROG	3.81	222.09	0.77	0.53
		REPG	3.34	233.26	0.85	0.52
	LDM	CFG (w=1.50)	3.60	246.67	**0.87**	0.48
	DiT-XL/2	CFG (w=1.50)	2.27	278.24	0.83	0.57
		PROG	2.25	279.36	0.82	0.58
		REPG	2.17	268.42	0.80	**0.60**
	VAR-$d30$	CFG (w=1.75)	1.93	315.64	0.82	0.59
		IGG (w=1.85)	**1.92**	**321.28**	0.82	0.59
512×512	ADM	CLSG	7.72	172.71	**0.87**	0.42
	DiT-XL/2	CFG	3.04	240.82	0.84	0.54
	VAR-$d36$	CFG (w=1.50)	2.61	293.7	0.82	0.56
		IGG (w=2.10)	**2.56**	**314.3**	0.82	**0.57**

room for improvement regarding the task of concentrating on semantically important tokens.

Qualitative results. We perform side-by-side visual comparisons between CFG and our method similar to our analysis in Section 4. Figure 2 depicts a representative comparison (additional examples can be found in Appendix C). In numerous cases, IGG gathers guidance signals towards semantically important tokens and forms contours clearly resembling the corresponding foreground objects, mimicking the patterns of CFG in diffusion modelling (Figure 1). This finding, along with the improvement of our method over CFG, further reinforces the notion that the superiority of diffusion models can be explained by the insights presented in Section 4.

6.2. Text-to-Image Generation

Quantitative results. To test the effectiveness of IGG on SwAR models for text-to-image tasks, we compare IGG against CFG on the MJHQ (Li et al., 2024a), MS-COCO (Lin et al., 2014), and GenEval (Ghosh et al., 2023) benchmarks. Two recent SwAR backbones, VAR-CLIP (Zhang et al., 2024) and Switti (Voronov et al., 2025), were considered for IGG, while we also include baselines for other popular backbones (SDXL (Podell et al., 2024), LlamaGen (Sun et al., 2024), and HART (Tang et al., 2025)) for ease of comparison. The metrics to compare were GenEval, FID, CLIP, PickScore (Kirstain et al., 2023), and ImageReward (IR) (Xu et al., 2023). As shown in Table 3, IGG helps Switti to achieve significant improvement on different metrics, especially Geneval and FID. These improvements translate to stronger complex prompt modelling capacity and arguably enhance image quality, as shown in the GenEval breakdown in Table 4. Specifically, significant performance gains were observed in Two-object or Colour attribution scenarios. The reasoning for these improvements is similar to class-conditioned generation, since the misalignment property is relaxed, images generated under IGG receive more guidance on semantically important regions without introducing more text-conditioned artefacts in less important regions.

Qualitative results Sample generations from Switti were shown in Figure 3 for comparison between the three guidance schemes: No guidance, CFG, and IGG. Observing the samples, we see that IGG achieves the best generation overall. With the no guidance scheme, due to receiving less text conditioning, the generation became too simple, with incomplete or prompt-disobedient objects. For CFG, while the objects were more complex, each patch receiving equal text-conditioning introduces artefacts to the generations. IGG, on the other hand, alternating guidance in regions of the image, resulted in less artefact-prone, more prompt-following and complex generations. This behaviour follows what we observed in the quantitative results, affirming the efficacy of our method.

6.3. Computational Cost

Table 5 shows a comparison between IGG and CFG in terms of inference time and training GFLOP count. The results indicate that, compared to CFG, **IGG incurs negligible computational overhead**, which is expected since IGG solely introduces an additional dot-product attention operation at each scale.

6.4. Metric Analysis

Guidance scale vs. evenness and divergence. To better understand the dynamics between guidance weight w and the evenness and divergence metrics detailed in Section 4, we computed these metrics over an extensive range of guidance weights on VAR-$d30$-IGG. However, we note that there is an inherent correlation between each metric with w since the nudges they evaluate include the guidance scales γ_k (see Equation 2), whose values depend on w by design. Thus, to disassociate these metrics, we scaled them by a factor corresponding to the (reciprocal of the) respective guidance scales applied. Both the original scores (dashed lines) and scaled scores (solid lines) were plotted in Figure 4 (left). Interestingly, evenness and divergence exhibit a similar trend: they improve with increasing w, at a decaying rate proportional to w. Notably, these two scores meet at $w = 1.35$ (black line). Evaluating VAR-$d30$-IGG

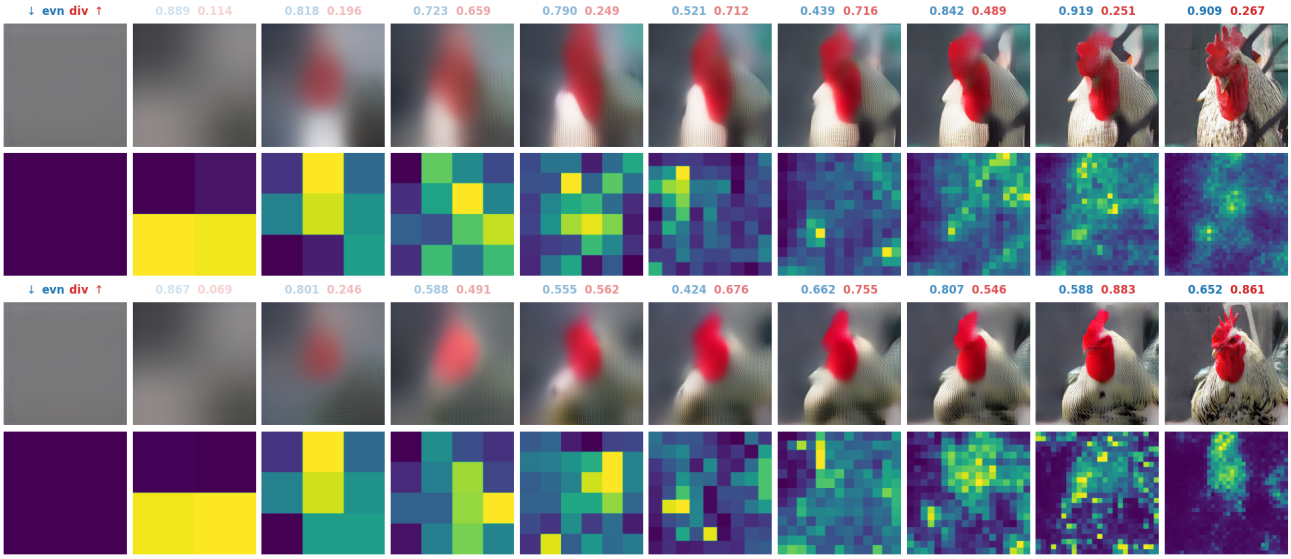


Figure 2. Comparison between CFG (top) and IGG(bottom) on ImageNet 512 × 512 class-conditioned generation (class: cock), with VAR (Tian et al., 2024) as the backbone model. Each column corresponds to a sampling step. Each heat map depicts the distribution of guidance on tokens at the respective sampling step, ranging from **purple** (weak guidance) to **yellow** (strong guidance). **Blue** and **red** scores indicate the evenness and divergence at each step, where ↑/↓ indicates that higher/lower is better (see Section 4 for further detail). **Our method improves upon classifier-free guidance by concentrating guidance towards regions of foreground objects.** Consequently, the hen in the IGG picture forms its face and body texture sooner, resulting in higher quality and fewer artefacts.

Table 3. Comparison of IGG against CFG on various text-to-image benchmarks using VAR-CLIP and Switti. Popular baselines using CFG are also included.↑/↓ indicates that higher/lower is better. Best results for each resolution are **bolded**. **Our method demonstrates strong overall performance across all benchmarks.**

Resolution	Model	Guidance	GenEval	MJHQ-30K				COCO-30K			
			Overall (↑)	FID (↓)	CLIP (↑)	PickScore (↑)	IR (↑)	FID (↓)	CLIP (↑)	PickScore (↑)	IR (↑)
256×256	VAR-CLIP	CFG	0.22	32.95	0.184	0.177	-1.78	10.95	**0.264**	**0.198**	**-0.87**
		IGG	**0.23**	**32.27**	**0.221**	**0.180**	**-1.58**	**10.93**	**0.264**	**0.198**	-0.89
512×512	SDXL	CFG	0.55	7.60	0.384	0.217	0.78	14.40	0.360	0.226	0.77
	LlamaGen	CFG	0.32	26.90	0.288	0.194	-0.45	44.80	0.274	0.208	-0.25
	HART	CFG	0.55	5.80	0.366	0.216	0.84	20.90	0.341	0.223	0.75
	Switti	CFG	0.62	9.50	0.388	**0.221**	**1.15**	17.60	0.355	**0.228**	0.93
		IGG	**0.64**	**7.09**	**0.389**	0.220	1.14	**16.90**	**0.357**	**0.228**	**0.97**

Table 4. Full evaluation breakdown of GenEval results. Best results for each model are **bolded**. **Our method improves prompt modelling capabilities.**

Model	Guidance	Single-object	Two-object	Counting	Colours	Position	Colour attribution
VAR-CLIP	CFG	0.694	0.068	**0.113**	0.431	0.003	0.003
	IGG	**0.716**	**0.078**	0.094	**0.452**	**0.008**	**0.005**
Switti	CFG	0.997	0.753	0.556	0.883	0.113	0.398
	IGG	**1.000**	**0.808**	**0.609**	**0.888**	**0.123**	**0.430**

at this guidance weight yielded an FID of ≈ 1.98, which is not too far off the optimal FID obtained at $w = 1.85$ (grey line). This result demonstrates the potential of evenness and divergence to approximate the optimal guidance weight to apply to a model, which could significantly alleviate the labour of hyperparameter tuning for large models.

Evenness and divergence vs. FID. We further investigated the relationship between the proposed metrics and FID, as depicted in Figure 4 (centre). The striking similarity between the (scaled) scores is once again clearly reflected. In particular, they both exhibit a remarkably sharp FID curve under varying conditions. The optimal FID is attained at almost-equal evenness and divergence (within 0.06 away). This strongly indicates a close underlying connection between evenness-divergence equilibrium and FID optimality, providing further support for the ability of these two metrics to approximate the optimal guidance weight.

Table 5. Inference time and GLOP count comparison. **Our method incurs negligible compute to the generation pipeline compared to CFG.**

Model	Guidance	Inference time (per image)	GFLOPs (per image)
VAR-$d30$	CFG	0.1391	259.93
	IGG	0.1408 (+0.0017)	259.93 (+0.00)
VAR-$d36$	CFG	0.7009	311.12
	IGG	0.7103 (+0.0004)	311.12 (+0.00)



Figure 3. Example 1024×1024 generations of Switti under three guidance schemes: no guidance, CFG, and IGG (ours). Without IGG, CFG or vanilla sampling of Switti has higher chance of generating failure features, around one in four samples.

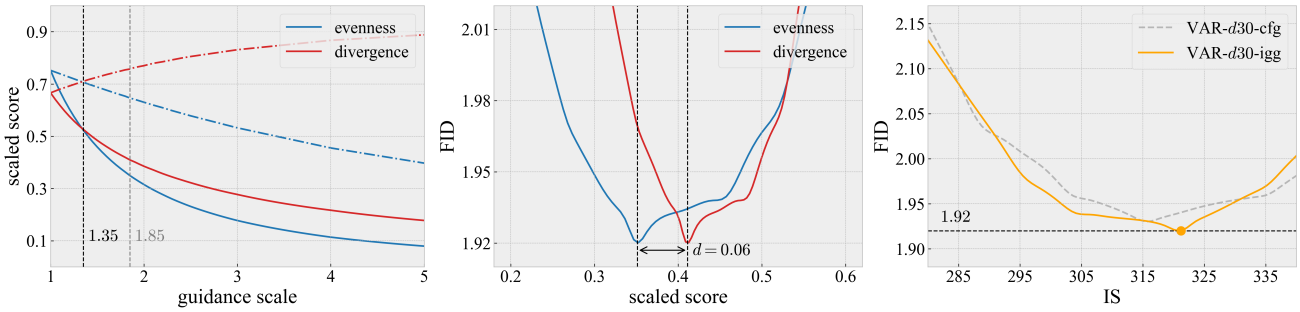


Figure 4. Analysing the relationships between various metric scores attained by VAR-$d30$-IGG. From left to right: effect of changing guidance scales on reported evenness and divergence scores, where dashed and solid lines depict raw and scaled scores, respectively; correspondence between FID and scaled evenness and divergence scores; and FID-IS trade-off curve.

Table 6. Ablation study on VAR-$d30$-IGG. Each row (except the first row) presents the results of a single modification to the original implementation (first row). ↑/↓ indicates that higher/lower is better. Best results are **bolded**.

Description	FID (↓)	IS (↑)	Pre (↑)	Rec (↑)	Evn (↓)	Div (↑)
Vanilla (*no changes*)	**1.92**	321.2	0.82	0.59	0.646	0.757
Mixed scheme (w_{IGG}=w_{CFG}=0.75)	5.48	**415.6**	**0.88**	0.48	**0.492**	**0.827**
Fixed schedule (γ_k=1.85)	4.29	359.5	0.87	**0.50**	0.727	0.684
Sliding window (w_k=$\sqrt{h_k w_k}$)	**1.92**	321.5	0.82	**0.60**	0.608	0.759

FID vs. IS. Figure 4 (right) depicts the trade-off between sample diversity (IS) and fidelity (FID) of our method compared to CFG. As IS increases, both methods initially achieve lower FID, but beyond IS of $\approx$ 320, further increases in diversity come at the cost of worsening fidelity, suggesting a natural limit to the achievable balance. Notably, within the proximity of the optimal trade-off, our method consistently outperforms CFG. Overall, the curve underscores that our method provides a slightly more favourable balance between diversity and realism. As suggested in Table 2, we expect this advantage to increase with model scale.

6.5. Ablation Study

Mixed guidance schemes. We investigated the potential of mixing (vanilla) CFG and IGG. In this experiment, we fixed an equal guidance weight w of 0.75 for both the CFG and IGG components. Although the resulting model was able to improve upon evenness and divergence, it fell short in terms of FID. This discrepancy emphasises the importance of achieving an equilibrium between evenness and divergence as opposed to arbitrarily improving them independently, as revealed in Section 6.4. It also, once again, calls for further investigation on the reason why mixing guidance schemes works so well in diffusion modelling.

Fixed guidance schedule. To assess the importance of choosing the right guidance schedule, we replaced the default schedule in Tian et al. (2024) with a fixed schedule. Specifically, we set the guidance scale γ_k at every scale level to 1.85. This change turns out to be detrimental to the model's performance, with a worse FID, evenness, and divergence compared to the vanilla implementation. This result is a testament to the major role that the ratio-based guidance schedule plays in the efficacy of IGG and, quite likely, also CFG.

Sliding-window guidance. The original implementation of IGG performs a global attention computation at each scale level (see Equation 6). Inspired by the successes of localised attention mechanism in NLP (Beltagy et al., 2020), we implemented a variant of IGG which utilises a *scale-wise 2-D sliding window* for attention computation. Compared to NLP-based sliding-window attention, there are two notable changes: (1) as token maps are inherently 2-D grids, the sliding window is consequently a 2-D sub-grid; (2) to account for the growth in scale at each sampling step, the size w_k of the sliding window is scaled accordingly at step k. In our experiment, we set $w_k := \sqrt{h_k w_k}$. The result of the experiment does not indicate any non-negligible improvement from the original implementation.

7. Conclusion

In this work, we investigated the underlying dynamics of guidance in scale-wise autoregressive (SwAR) models and, via the introduction of the evenness and divergence metrics, revealed a fundamental limitation: unlike diffusion models, guidance in SwAR models is often dispersed and misaligned with respect to visual semantics. Building upon this key insight, we introduced IGG, a novel guidance framework that explicitly aims to anchor guidance signals to semantically important tokens via an adaptive weighting function, which we realised using the self-attention mechanism. Our extensive experiments across both class-conditioned and text-to-image generation tasks successfully demonstrated that IGG consistently improves structural fidelity, coherence, and prompt alignment over baseline methods. Overall, this works promotes the core principle that not all tokens should be guided equally, opening promising new avenues for future research on innovative guidance strategies in SwAR generative modelling by utilising our proposed framework.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer, December 2020. arXiv:2004.05150 [cs].

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. MaskGIT: Masked Generative Image Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315, 2022. doi: 10.1109/CVPR52688.2022.01103.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Dinh, A.-D., Liu, D., and Xu, C. PixelAsParam: A Gradient View on Diffusion Sampling with Guidance. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8120–8137. PMLR, July 2023a. ISSN: 2640-3498.

Dinh, A.-D., Liu, D., and Xu, C. Rethinking conditional diffusion sampling with progressive guidance. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 42285–42297. Curran Associates, Inc., 2023b.

Dinh, A.-D., Liu, D., and Xu, C. Representative Guidance: Diffusion Model Sampling with Coherence. In *The Thirteenth International Conference on Learning Representations*, 2025.

Esser, P., Rombach, R., and Ommer, B. Taming Transformers for High-Resolution Image Synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878. IEEE, 2021. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.01268.

Fan, L., Li, T., Qin, S., Li, Y., Sun, C., Rubinstein, M., Sun, D., He, K., and Tian, Y. Fluid: Scaling Autoregressive Text-to-image Generative Models with Continuous Tokens. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ghosh, D., Hajishirzi, H., and Schmidt, L. GenEval: An object-focused framework for evaluating text-to-image alignment. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023.

Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., and Liu, X. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744. IEEE, 2025.

Ho, J. and Salimans, T. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Hong, S., Lee, G., Jang, W., and Kim, S. Improving Sample Quality of Diffusion Models Using Self-Attention Guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7428–7437, 2023. doi: 10.1109/ICCV51070.2023.00686.

Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., and Laine, S. Guiding a Diffusion Model with a Bad Version of Itself. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024a.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and Improving the Training Dynamics of Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, Seattle, WA, USA, June 2024b. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02282.

Khanam, R. and Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements, October 2024. arXiv:2410.17725 [cs].

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and

Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023.

Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., and Lehtinen, J. Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122458–122483. Curran Associates, Inc., 2024.

Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., and Doshi, S. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation, February 2024a. arXiv:2402.17245 [cs].

Li, P., Yan, S., Tsai, J., Zhang, R., An, R., Guo, Z., and Gao, X. Adaptive Classifier-Free Guidance via Dynamic Low-Confidence Masking, May 2025. arXiv:2505.20199 [cs].

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive Image Generation without Vector Quantization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 56424–56445. Curran Associates, Inc., 2024b.

Liao, B., Xu, Y., Dong, H., Li, J., Monz, C., Savarese, S., Sahoo, D., and Xiong, C. Reward-Guided Speculative Decoding for Efficient LLM Reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Liu, H., Jiang, B., Xiao, Y., and Yang, C. Coherent Semantic Attention for Image Inpainting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4169–4178, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00427.

Nisonoff, H., Xiong, J., Allenspach, S., and Listgarten, J. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peebles, W. and Xie, S. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Pielou, E. C. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144, December 1966. ISSN 0022-5193. doi: 10.1016/0022-5193(66)90013-0.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*, October 2024.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ren, S., Yu, Q., He, J., Shen, X., Yuille, A., and Chen, L.-C. FlowAR: Scale-wise Autoregressive Image Generation Meets Flow Matching, December 2024. arXiv:2412.15205 [cs].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., and Angelova, A. TokenLearner: What Can 8 Learned Tokens Do for Images and Videos? In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Schiff, Y., Sahoo, S. S., Phung, H., Wang, G., Boshar, S., Dalla-torre, H., Almeida, B. P. d., Rush, A., Pierrot, T., and Kuleshov, V. Simple Guidance Mechanisms for Discrete Diffusion Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., and Yuan, Z. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation, June 2024. arXiv:2406.06525 [cs].

Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., and Han, S. HART: Efficient Visual Generation with Hybrid Autoregressive Transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865. Curran Associates, Inc., 2024.

Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., and Baranchuk, D. Switti: Designing Scale-Wise Transformers for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2025.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. Image Captioning With Semantic Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, 2016.

Zhang, H., Zhang, J., and Wan, X. How Much To Guide: Revisiting Adaptive Guidance in Classifier-Free Guidance Text-to-Vision Diffusion Models, June 2025. arXiv:2506.08351 [cs].

Zhang, Q., Dai, X., Yang, N., An, X., Feng, Z., and Ren, X. VAR-CLIP: Text-to-Image Generator with Visual Auto-Regressive Modeling, 2024.

Zheng, G., Li, S., Wang, H., Yao, T., Chen, Y., Ding, S., and Li, X. Entropy-Driven Sampling and Training Scheme for Conditional Diffusion Generation. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, volume 13682, pp. 754–769. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-20046-5 978-3-031-20047-2. doi: 10.1007/978-3-031-20047-2_43. Series Title: Lecture Notes in Computer Science.

A. Computing the Divergence Score

Algorithm 1 details the procedure for systematically computing the divergence score as defined in Section 4.

Algorithm 1 Procedure for computing divergence score.

Require: Sampled image x and associated token maps $s := (s_1, \cdots, s_K)$ of sizes $\{h_k, w_k\}_{k=1}^{K}$.
Ensure: Divergence score for x.

$M :=$ segment(x) {Obtain binary segmentation mask}
$J := 0$
for $k \in 2..K$ {Skip s_1 (see Section 4)} **do**
  $M_k :=$ interpolate(M, h_k, w_k) {Down-sample mask to $h_k \times w_k$}
  $p_k^{\rightarrow} := p_\theta^{\rightarrow}(s_k|c)$ {Obtain guided nudge distribution}
  $s_k^* :=$ sample$(s_k[\neg M_k], h_k, w_k)$ {Sample with replacement $h_k w_k$ unguided tokens}
  $q_k^{\rightarrow} := p_\theta^{\rightarrow}(s_k^*|c)$ {Obtain unguided nudge distribution}
  $J := J + h_k w_k / \sum_{i=2}^{K} h_i w_i \cdot$ divergence$(p_k^{\rightarrow}, q_k^{\rightarrow})$ {Weighted mean analogous to PEI}
end for
return J

B. Comparing Guidance in Diffusion and SwAR Models

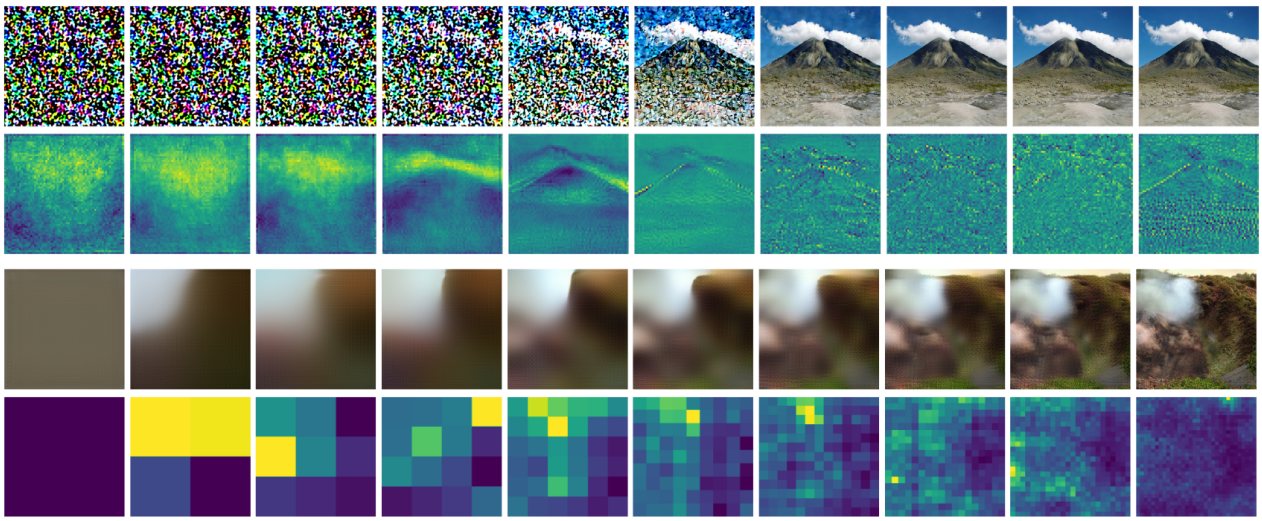
Figure 5 presents additional side-by-side comparisons of the guidance signals observed during the sampling processes of EDM2 (Karras et al., 2024b) and VAR (Tian et al., 2024). These visualisations make it easier to contrast how guidance is distributed in diffusion models compared to SwAR models. In particular, EDM2 exhibits relatively stable and balanced guidance, while VAR tends to experience progressively weaker signals across sampling steps. This distinction provides further evidence for our hypothesis that the uneven distribution of guidance contributes to the sampling quality gap between the two model families.

C. Comparing CFG and IGG

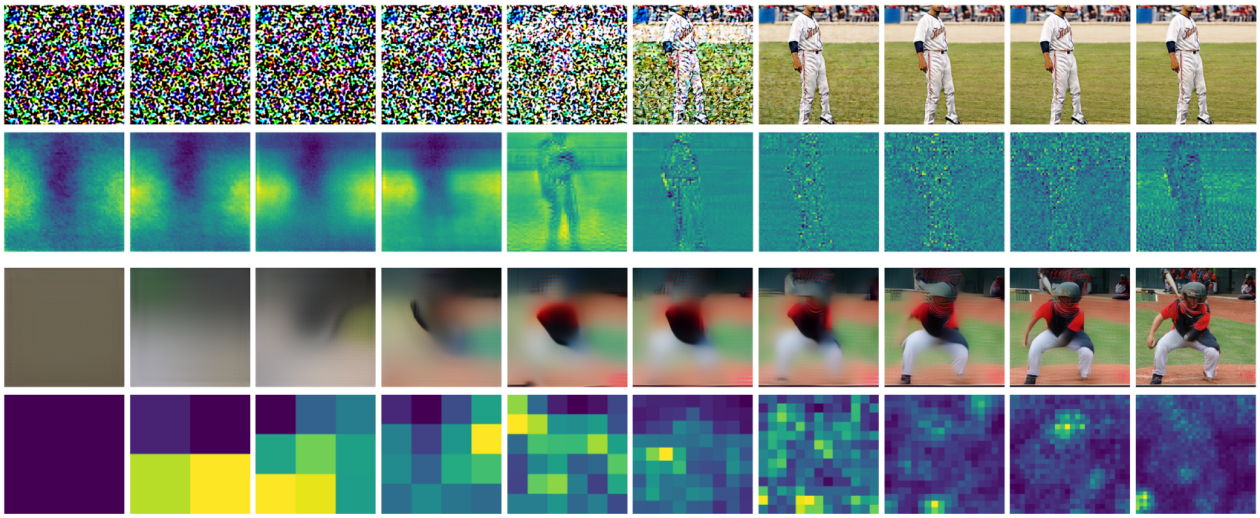
In Figure 6, we present additional results comparing samples generated with CFG against those obtained using IGG. These examples highlight the qualitative differences between the two approaches, particularly in terms of semantic coherence and visual fidelity. While CFG often struggles to maintain consistency across fine-grained details, IGG demonstrates a stronger ability to emphasise and preserve semantically important features. These examples further illustrate how our method balances guidance strength without sacrificing diversity in the generated outputs.

D. Guidance in Diffusion Models Across All Timesteps

Figure 7 visualises the full sampling process of EDM2 (Karras et al., 2024b). Interestingly, guidance signals seem to be the most pronounced in the middle of the sampling process, suggesting that the role played by guidance is the most influential around this period. This is consistent with the finding of Kynkäänniemi et al. (2024), providing further support for our proposed interpretation of CFG.

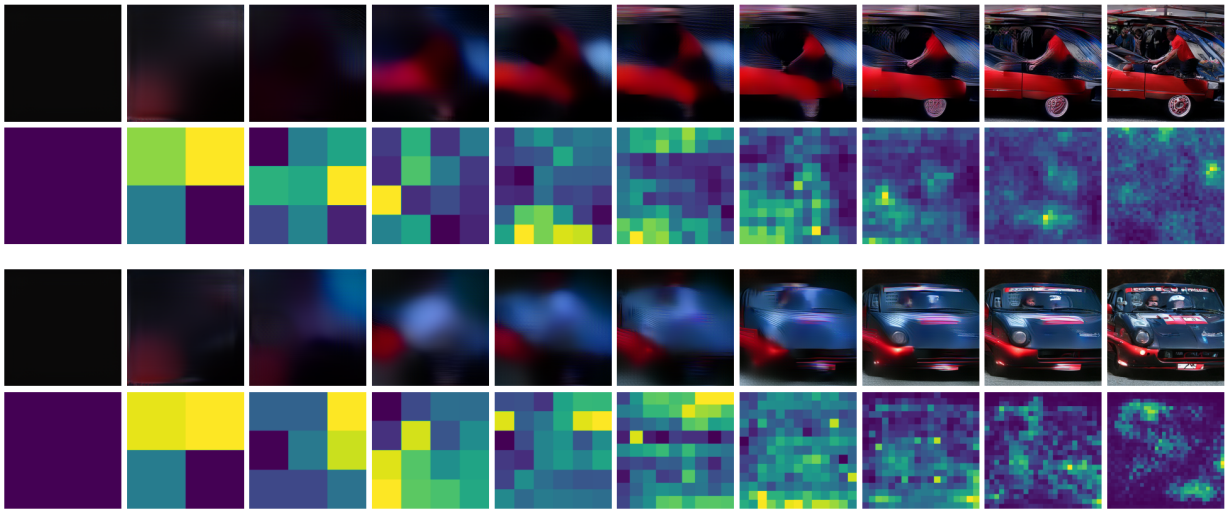


(a) volcano

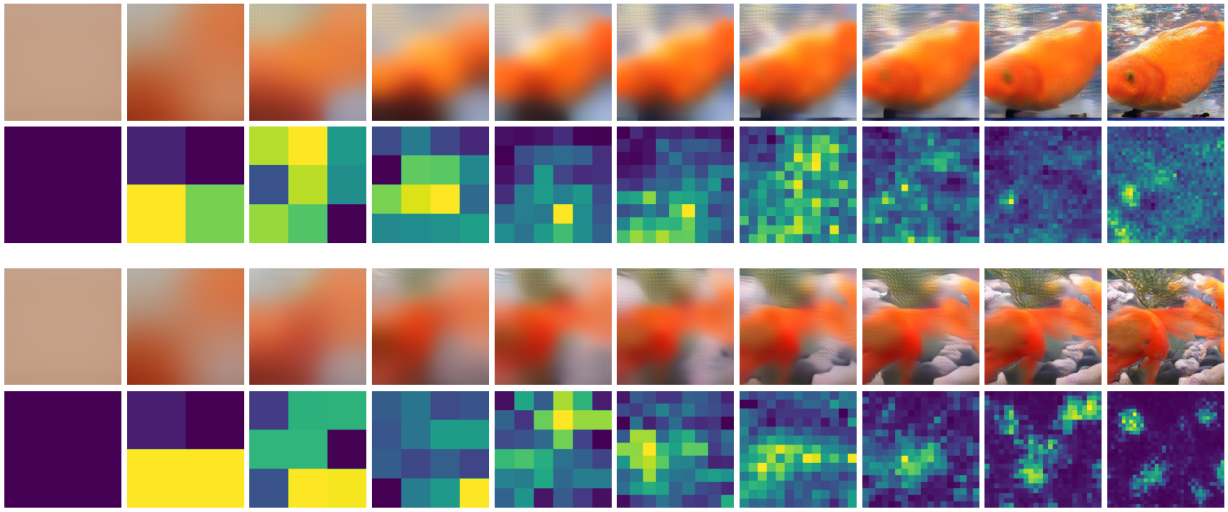


(b) baseball player

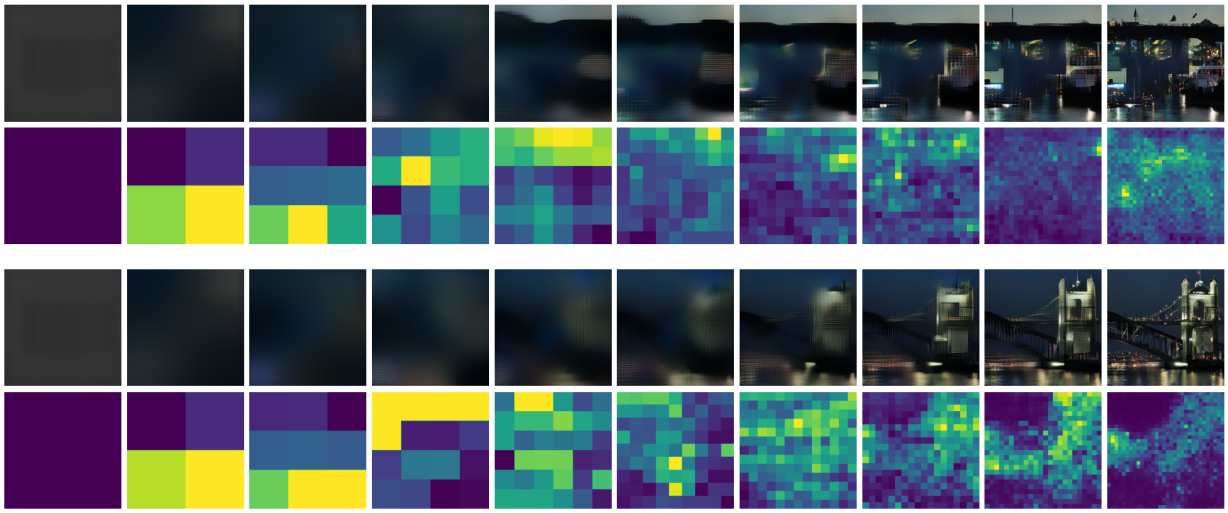
Figure 5. Additional comparisons of guidance signals in EDM2 (Karras et al., 2024b) (top) and VAR (Tian et al., 2024) (bottom).



(a) sports car



(b) goldfish



(c) pier

Figure 6. Additional comparisons of sampling in VAR-*d*36 (Tian et al., 2024) using CFG and IGG.

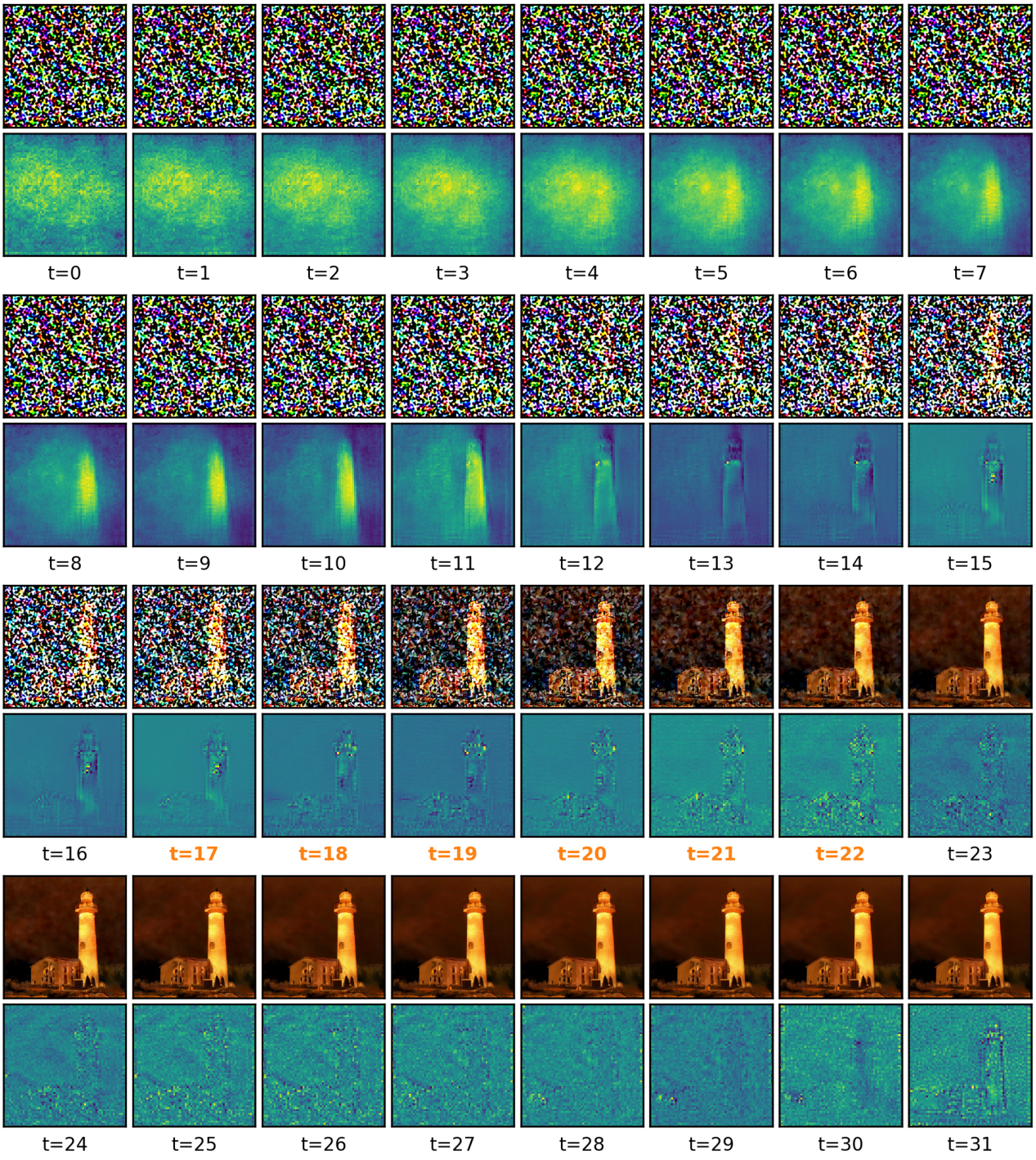


Figure 7. Visualisation of guidance signals for every timestep in the original sampling process of EDM2 (Karras et al., 2024b). Timesteps falling within the guidance interval reported in Kynkäänniemi et al. (2024) are highlighted in orange. This interval coincides with the transition from noise to image, which is where guidance plays the most influential role.