Guiding Visual Autoregressive Models through Spectrum Weakening

Chaoyang Wang **Tianmeng Yang** **Yunhai Tong**
State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
cywang@stu.pku.edu.cn

Abstract

Classifier-free guidance (CFG) has become a widely adopted practice for enhancing generation quality and improving condition alignment. Recent studies have explored guidance mechanisms for unconditional generation, yet these approaches remain fundamentally tied to assumptions specific to diffusion models. In this work, we propose a spectrum weakening framework for visual autoregressive (AR) models. The method is training-free and condition-free, constructing a controllable weak model in the spectral domain without architectural modifications. Specifically, we selectively retain spectral components along the channel dimension of internal representations to weaken features without requiring a complete spatial layout. We further apply spectrum renormalization to maintain feature scales during generation. Comprehensive empirical and ablation studies confirm the effectiveness of our approach, including raster-scan, random-order, and scale-wise AR, spanning discrete and continuous modeling and covering class and text-condition scenarios, demonstrating high-quality unconditional generation and strong prompt alignment for conditional generation. Project Page: https://wang-chaoyang.github.io/project/swg.

1 Introduction

Autoregressive (AR) modeling has recently emerged as a promising paradigm in visual generation, demonstrating remarkable scalability and controllability across a wide range of downstream applications, including video synthesis [1, 5], artistic creation [45], and high-fidelity image generation [24]. Unlike diffusion models [43, 44, 14] that iteratively refine noisy images as a whole, visual AR models construct images token by token, predicting each part conditioned on previously generated content.

Despite their success, existing visual generation models [10, 9, 21, 34, 33] often depend on classifier-free guidance (CFG) [15], which requires learning unconditional predictions alongside conditional generation. However, this requirement constrains the training setup, limits applicability to models without unconditional training, and does not directly provide guidance for unconditional generation. Recently, inspired by contrastive decoding [25], several works have proposed guiding the generation process using a weaker model [20] rather than an explicit classifier. These methods have shown strong performance in diffusion models, typically by modifying network architectures [19], designing perturbed timesteps [38], or perturbing self-attention maps [16, 17, 2]. However, as shown in Tab. 1, these strategies are challenging to transfer to general AR generation, which has no notion of diffusion timesteps and does not form a coherent global image representation until the entire sequence is decoded.

In light of these challenges, we propose a training-free spectrum weakening guidance (SWG) approach for visual AR generation. Specifically, our method has four key properties: 1) no reliance on external conditions, naturally supporting unconditional generation; 2) preservation of the original architecture, with no structural changes or retraining; 3) no assumption of a complete global layout, fitting AR

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



Figure 1: Visual results of Spectrum Weakening Guidance (SWG) on Lumina-mGPT. (a) The model synthesizes images using a null prompt. (b) The model synthesizes images using the specified prompts, shown as white text in the first row.

decoding with only partial context; and 4) enabling flexible adjustment of guidance intensity. As shown in Fig. 1, our proposed SWG significantly improves visual quality in both conditional and unconditional generation.

Concretely, we apply the discrete Fourier transform (DFT) along the channel dimension of intermediate features and use a binary mask to selectively retain spectral components. This provides a flexible way to construct weakened representations while preserving their shape and token layout. Since spectrum selection can alter feature magnitudes, we apply a renormalization step to maintain feature scales during generation. Empirically, we broadly compare and validate the effectiveness of our SWG, including raster-scan, random-order, and scale-wise AR, spanning discrete and continuous modeling and covering class and text-condition scenarios, and observe consistent quality gains. Further ablation studies demonstrate that the constructed weak model produces more uncertain predictions, and outperforms alternative weak-branch constructions.

The main contributions are as follows: 1) In this paper, we investigate and propose a training-free and condition-free weak model guidance approach for visual AR generation. 2) We design a spectrum selection pipeline to construct weakened intermediate representations, with renormalization to maintain feature scales during generation. 3) Comprehensive comparative and ablation experiments demonstrate the encouraging effectiveness and generalization of our method, as well as flexible controllability and compatibility with existing guiding techniques.

2 Related Work

Visual Autoregressive Model is motivated by the success of large language models (LLMs) [3, 47] and has recently emerged as a promising alternative to diffusion models because of its strong scalability [46]. Unlike diffusion models [27, 29, 30, 34, 37, 43, 44] that iteratively refine a complete image from noise, AR models generate visual tokens sequentially, conditioning each prediction on previously generated tokens. Existing approaches can be broadly divided into raster-scan models [28, 45, 36, 7, 8, 53, 50, 51, 22, 31], which decode images in spatial order, and set- or scale-wise models [24, 5, 4, 18, 54, 48], such as masked or hierarchical AR variants. Despite their success, current AR frameworks still rely heavily on CFG and remain limited in unconditional generation. In this work, we revisit this limitation through the lens of weak model guidance and introduce a training-free spectrum-weakening strategy tailored for AR generation.

Weak Model Guidance has recently attracted growing attention as an alternative to CFG, offering a more flexible and interpretable way to control generative models. Existing approaches to weakening model alignment can be broadly grouped by their underlying mechanisms. Some methods derive guidance by leveraging alternative training schedules or conditioning mechanisms. For example, Autoguide [20] uses states at varying convergence levels, while CFG [15] relies on jointly learned unconditional predictions. However, these methods require modified training procedures and precise computational control, and CFG is inherently limited to conditional setups, as the guidance process relies entirely on a conditioning signal. Others [19] adjust the model structure by layer pruning.

Table 1: Comparison of training-free self-guidance methods across generation paradigms. Existing approaches are designed for diffusion models or limited to scale-wise AR. Our proposed SWG supports all three AR paradigms.

	Diffusion models						*Visual autoregressive*	
	SAG [17]	PAG [2]	SEG [16]	TSG [38]	STG [19]	TPG [35]	SSG [41]	SWG (Ours)
Raster-scan AR	×	×	×	×	×	×	×	✓
Random-order AR	×	×	×	×	×	×	×	✓
Scale-wise AR	×	×	×	×	×	×	✓	✓

In addition, several methods exploit characteristics unique to diffusion models, such as perturbing self-attention maps spatially [2, 17, 16], shuffling image tokens [35], or perturbing timesteps [38]. However, these designs rely on the full-image generation and timestep formulations specific to diffusion models, making them unsuitable for AR generation.

In this work, we revisit alignment weakening from an information-reduction perspective, applying spectrum selection to attenuate model alignment without modifying the architecture.

Spectral Representations offer a powerful perspective for processing visual signals in the frequency domain. The transformation between spatial and spectral domains, typically implemented via the DFT, enables bidirectional conversion without significant information degradation. In the spectral domain, low-frequency components capture coarse structures and overall geometry, while high-frequency components encode fine details and textures, providing an interpretable decomposition of visual information [49]. This property has been exploited in various diffusion-based generation and control frameworks [42] to enhance structure–detail balance. Beyond interpretability, spectral representations also enable the decoupling of distinct signal components, inspiring explorations in AR modeling [52]. In contrast, we operate in the spectral domain for flexible information reduction. Spectral selection suppresses selected components while preserving feature shapes and token positions, providing a controllable alternative to spatial perturbations.

3 Method

3.1 Preliminary

Visual Autoregressive Model. In an AR framework, the conditional distribution of an image x given a condition c is factorized as:

$$p(x|c) = \prod_{t=1}^{T} p(x_t|x_{<t}, c), \tag{1}$$

where x_t denotes the t-th token in the image sequence. The model is optimized by maximizing the likelihood.

To incorporate conditioning signals, such as text prompts, the model encodes the corresponding tokens and concatenates them with image tokens, enabling it to attend to both modalities jointly. During inference, the KV-cache stores the key-value representations of previously generated tokens, enabling efficient computation as the query Q attends to the current step while K and V aggregate information from both past and current tokens.

CFG is also commonly used in visual AR models. An unconditional branch is trained with null prompts, and during generation, the logits from the conditional and unconditional branches are linearly combined to balance fidelity and diversity.

Spatial to Spectral Transformation. Let $\mathbf{x} \in \mathbb{C}^N$ denote a discrete spatial-domain signal and let $\hat{\mathbf{x}} \in \mathbb{C}^N$ denote its discrete Fourier transform spectrum. The discrete Fourier transform can be written in matrix form as

$$\hat{\mathbf{x}} \;=\; \mathbf{W}\,\mathbf{x}, \tag{2}$$

where the DFT matrix $\mathbf{W} \in \mathbb{C}^{N\times N}$ has entries

$$W_{k,n} \;=\; \omega^{kn}/\sqrt{N}, \qquad k, n = 0, 1, \ldots, N-1, \tag{3}$$

and

$$\omega \triangleq e^{-\mathrm{i}\,2\pi/N} \tag{4}$$

is the primitive N-th root of unity (here $\mathrm{i} = \sqrt{-1}$).

The inverse discrete Fourier transform (IDFT) under the common normalization convention is

$$\mathbf{x} = \mathbf{W}^* \hat{\mathbf{x}}, \tag{5}$$

where $\mathbf{W}^*$ denotes the conjugate transpose of $\mathbf{W}$.

Specifically, the DFT matrix $\mathbf{W}$ satisfies $\mathbf{W}^* \mathbf{W} = \mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix.

3.2 Spectral Weakening Guidance

To derive a weaker variant of the original model, we attenuate the information content of its internal representations via spectral selection. We first outline the overall pipeline and then discuss the effect of spectral selection from an information-theoretic perspective. Finally, we introduce a spectrum renormalization step to maintain feature scales and improve numerical stability during generation.

3.2.1 Selective Spectrum Preservation

We introduce a spectrum selection pipeline to modulate the information within visual AR models. Given feature tensor $\mathbf{x} \in \mathbb{R}^C$, we apply DFT along the channel dimension to obtain its spectral representation $\hat{\mathbf{x}} = \mathbf{W}\,\mathbf{x}$ following Eq. 2. We employ a binary spectral mask to select a subset of spectral components, which is equivalently represented as a diagonal selection matrix

$$\mathbf{M} = \mathrm{diag}(M_0, M_1, \ldots, M_{C-1}) \in \{0,1\}^{C \times C}, \tag{6}$$

where $M_i = 1$ indicates that the i-th frequency is preserved, and $M_i = 0$ indicates suppression.

The filtered spectrum and its inverse transform are given by

$$\hat{\mathbf{x}}' = \mathbf{M}\hat{\mathbf{x}}, \qquad \mathbf{x}' = \mathbf{W}^* \hat{\mathbf{x}}'. \tag{7}$$

Since $\mathbf{x}$ is real-valued, its Fourier spectrum $\hat{\mathbf{x}}$ exhibits conjugate symmetry, and numerical computation may introduce imaginary residuals. Therefore, we take the real part of the reconstructed signal as the final output:

$$\mathbf{x}' \leftarrow \Re(\mathbf{x}'). \tag{8}$$

The resulting $\mathbf{x}'$ retains the same shape as $\mathbf{x}$ while suppressing selected spectral components. Spectral selection acts along the channel dimension through a fixed binary mask, yielding a linear and differentiable transformation of the input and allowing controllable feature perturbations while preserving token positions.

Discussion. The proposed spectrum selection pipeline is flexible, as it does not rely on the availability of the complete image structure during generation. Furthermore, contemporary AR models typically rely on the KV cache for inference, in which each block has access to only one visible token. Consequently, structural perturbations such as patch shuffling [35] are infeasible in this setting.

3.2.2 Information Reduction Analysis

Spectral selection reduces the information available to subsequent layers by suppressing selected frequency components. Since the DFT is invertible, transforming features into the spectral domain does not alter their information content. In contrast, applying a binary spectral mask discards selected coefficients.

Formally, let $\mathbf{Z}$ be any random variable potentially dependent on $\mathbf{x}$. Since the spectrally filtered representation $\mathbf{x}'$ is a deterministic function of $\mathbf{x}$, the data processing inequality gives

$$\mathsf{I}(\mathbf{x}'; \mathbf{Z}) \leq \mathsf{I}(\mathbf{x}; \mathbf{Z}). \tag{9}$$

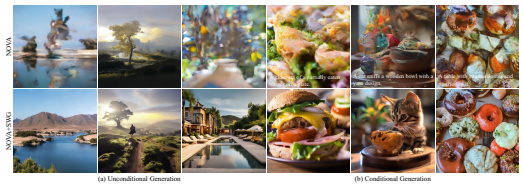
This indicates that spectral selection cannot increase the information about $\mathbf{Z}$ contained in the original representation. By controlling the preserved spectral components, we obtain a simple mechanism for constructing weakened representations for guidance.

Algorithm 1: Sampling with Spectrum Weak Guidance

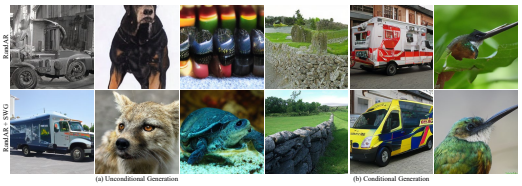
Input: SWG scale ω_s; CFG scale ω_c (opt.); binary spectrum mask $\mathbf{M}$; condition c; SWG module set $\mathcal{S} \subseteq \theta = \{\theta_1, \ldots, \theta_N\}$

Output: Generated sample x_T

$x_0 \leftarrow \langle \text{BOS} \rangle$
for $t = 1, \ldots, T$ **do**
  $z_{t,c} \leftarrow f_\theta(x_{t-1}, c)$ ▷ `Cond. pred.`
  $\ell_0 \leftarrow x_{t-1}$
  for $i = 1, \ldots, N$ **do**
    $\ell_i \leftarrow f_{\theta_i}(\ell_{i-1}, c)$
    if $\theta_i \in \mathcal{S}$ **then**
      $\hat{\ell}_i \leftarrow \mathbf{M}\mathbf{W}\ell_i$
      $\ell_i \leftarrow \text{Renorm}(\hat{\ell}_i, \ell_i)$
  $z_{t,p} \leftarrow \ell_N$; $z_t \leftarrow z_{t,c} + \omega_s(z_{t,c} - z_{t,p})$ ▷ `Apply SWG`
  if *use CFG* **then** ▷ `Optional CFG`
    $z_{t,b} \leftarrow f_\theta(x_{t-1}, \emptyset)$; $z_t \leftarrow z_t + \omega_c(z_{t,c} - z_{t,b})$
  $x_t \leftarrow \text{Sample}(\text{Softmax}(z_t))$
return x_T



(a) NOVA. The left three columns show null-prompt generations; the right three columns show prompt-conditioned ones. Spatial renormalization is used.



(b) RandAR. The left three columns show unconditional generations; the right three show conditional ones.

Figure 2: Visual results of SWG on two autoregressive models.

3.2.3 Spectrum Renormalization

Spectral selection may alter the magnitude of intermediate representations. We therefore apply a renormalization step to maintain feature scales during generation.

Specifically, given the original spectrum $\hat{\mathbf{x}} = \mathbf{W}\mathbf{x}$ and the filtered spectrum $\hat{\mathbf{x}}' = \mathbf{M}\hat{\mathbf{x}}$, the renormalized representation can be obtained in the spectral domain as

$$\tilde{\hat{\mathbf{x}}}' = \hat{\mathbf{x}}' \cdot \frac{\|\hat{\mathbf{x}}\|_2}{\|\hat{\mathbf{x}}'\|_2 + \epsilon}, \quad \tilde{\mathbf{x}}' = \Re\big(\mathbf{W}^* \tilde{\hat{\mathbf{x}}}'\big), \tag{10}$$

or alternatively, after reconstruction in the spatial domain as

$$\tilde{\mathbf{x}}' = \mathbf{x}' \cdot \frac{\|\mathbf{x}\|_2}{\|\mathbf{x}'\|_2 + \epsilon}. \tag{11}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm, and $\epsilon > 0$ is a small constant for numerical stability.

We refer to these two implementations as *spectral renormalization* and *spatial renormalization*, respectively. For simplicity, we collectively denote this step as Renorm in Alg. 1.

4 Experiment

4.1 Experimental Settings

Baselines. Our study encompasses random-order, raster-scan, and scale-wise AR paradigms, as shown in Tab. 3. Since most prior methods rely on the characteristics of diffusion or require access to

Table 2: Performance of different visual AR models on the COCO dataset. Both unconditional and conditional generation are considered. CS and AS denote CLIP score and aesthetic score, respectively.

(a) Unconditional generation

Method	FID ↓	IS ↑	AS ↑
NOVA	85.34	12.25	4.03
+SSG [41]	-	-	-
+SWG (Ours)	50.37	19.12	5.75
Lumina-mGPT	103.53	9.71	5.82
+SSG [41]	-	-	-
+SWG (Ours)	101.46	9.33	6.85
Infinity	142.61	5.45	5.52
+SSG [41]	157.27	4.76	5.65
+SWG (Ours)	133.45	6.42	5.75

(b) Conditional generation

Guidance	FID ↓	IS ↑	CS ↑	AS ↑
NOVA	46.86	19.46	26.62	4.21
+SSG [41]	-	-	-	-
+SWG (Ours)	15.80	31.27	29.86	5.78
Lumina-mGPT	37.33	24.50	28.07	5.98
+SSG [41]	-	-	-	-
+SWG (Ours)	29.91	28.43	29.28	6.70
Infinity	33.92	35.43	31.21	6.23
+SSG [41]	31.49	35.84	31.22	6.24
+SWG (Ours)	30.50	38.76	31.49	6.34

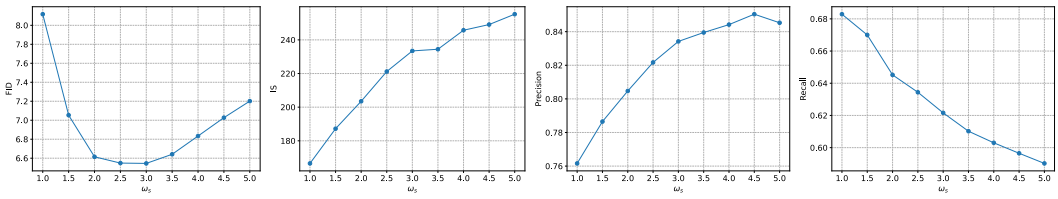


Figure 3: Quantitative evaluation of RandAR under different SWG guidance strength ω_s. From left to right, the curves depict FID, IS, Precision, and Recall with respect to the guidance scale.

global spatial structure at every inference step, we include SSG in our comparison of scale-wise AR models. Its original formulation does not directly extend to the other two AR paradigms.

Table 3: Baseline Model Illustration. Token prediction types: Cont. (continuous) and Disc. (discrete). AR inference paradigms: Rand. (random-order), Rast. (raster-scan), and Scale. (scale-wise).

Model	Input	Resolution	#Params	Token Type	AR Type
RandAR [32]	Class	256	0.7B	Disc.	Rand.
Lumina-mGPT [28]	Text	768	7B	Disc.	Rast.
NOVA [5]	Text	1024	0.3B	Cont.	Rand.
Infinity [11]	Text	1024	2B	Disc.	Scale.

Implementation Details. By default, we retain 25% of spectral components, using the same masking strategy across all modified layers. Unless otherwise specified, we apply SWG to layers 4-7 for Lumina-mGPT, RandAR, and Infinity, and to layers 8-11 for NOVA. Experiment-specific settings are provided in the appendix. The CFG scale (if used) follows their official implementations.

Datasets and Metrics. Following common practice [16, 38, 35], we employ COCO dataset [26] for text-to-image task evaluation, including both the conditional and unconditional generation settings. For class-to-image tasks, we use ImageNet dataset [6] for evaluation. Quantitative results are reported using Fréchet inception distance (FID) [13], Inception Score (IS) [39], CLIP score [12], Aesthetic score [40], precision and recall [23].

4.2 Main Results

Qualitative Results. We visually illustrate the effectiveness of the proposed SWG in improving the synthesized results. As shown in Fig. 1 and 2, these models produce blurred or distorted images without guidance. Moreover, the generated results do not align correctly with the prompts; e.g., 'heron' is misinterpreted as 'woman' in the 6^{th} column in Fig. 1. On the contrary, our proposed SWG significantly improves the quality of synthesized results on both discrete and continuous AR, as well as on unconditional and conditional generation, demonstrating its effectiveness and generalization.

Note that in unconditional generation scenarios, the reference model itself is the base model, therefore the standard CFG is not applicable.

We also evaluate SWG on class-condition generation. As shown in Fig. 2b, SWG significantly improves the quality of synthesized results, improving object deformation and blurry outlines.

Additionally, we examine the relationship between the guided and baseline results. Interestingly, the synthesized pairs from NOVA exhibit higher spatial relevance, whereas the pairs from Lumina-mGPT appear to differ. We hypothesize that this phenomenon is attributed to a more compact sampling space of discrete AR models.

Quantitative Comparisons. Our proposed SWG is competent for both unconditional and conditional scenarios on text-to-image generation tasks. As shown in Tab. 2a and Tab. 2b, SWG improves FID and aesthetic scores across different AR modeling strategies. For the continuous token prediction model NOVA, the gains are particularly pronounced: FID is reduced by 34.97 in unconditional generation and by 31.06 in conditional generation, with simultaneous improvements in terms of text alignment and aesthetics. For discrete token prediction models, Lumina-mGPT and Infinity, the improvements are more moderate but still consistent. SWG steadily lowers FID and enhances aesthetics in both settings. Specifically, Lumina-mGPT improves CLIP score by 1.21, indicating a stronger ability to follow textual instructions.

Fig. 3 shows the effectiveness of our SWG on the class-conditioned generation task. As the guidance strength ω_s increases, the FID score first decreases and then rises, reaching the best value of 6.55 at $\omega_s = 3.0$, indicating an optimal balance between fidelity and diversity. The Inception score increases steadily from 166.56 at $\omega_s = 1$ to 255.27 at $\omega_s = 5$, reflecting consistent improvements in perceptual quality. Precision also tends to improve overall from 0.76 to around 0.85. Meanwhile, recall gradually decreases from 0.68 to 0.59, suggesting that stronger guidance increases fidelity at the expense of reduced diversity. These results indicate that moderate SWG strength yields the most balanced and visually coherent generation quality.

4.3 Ablation Studies

Spectrum Analysis. We consider two key factors regarding spectral bands in weak model construction: bandwidth and location. To achieve this, we perform fine-grained sweep ablation. Specifically, we sort the DFT-transformed spectrum according to harmonics, with the entire range denoted as 0-1.0. Then, we set start and end endpoints with a step size of 0.2. Excluding invalid cases where the end point is earlier than or equal to the start point, we obtain a total of 15 images, arranged in a pyramid pattern, as shown in Fig. 4 (a).

Interestingly, a pattern seems to emerge. The generated results gradually become clearer along the pyramid from the lower left to the upper right (equivalent to from left to right, from bottom to top). However, the generated details seem to decrease (the room becomes neater), suggesting that our *channel-wise perturbation* is fundamentally different from previous *spatial-wise filtering* methods.

We attribute this phenomenon to the reduction in information. Due to the varying prediction difficulty across tokens and the cumulative effect of errors, we examine entropy differences to compare the uncertainty of the two branches. We ablate the bandwidth and position separately, with results shown in Fig. 4 (b) and (c), respectively. On the one hand, changing the bandwidth has a pronounced effect on the entropy difference. Specifically, as the bandwidth increases from 0.2 to 0.8, the entropy gap between the weak and base branches generally becomes smaller at corresponding inference steps. This trend is consistent with the less pronounced guidance effects observed in the corresponding visual examples. On the other hand, if the bandwidth is kept constant, changing only the location of the retained frequency band has a comparatively smaller effect on the entropy difference, and the curves exhibit broadly similar trends throughout generation.

Ablation of Renormalization. As shown in Fig. 5, different signal renormalization strategies yield varying synthesis results. Applying renormalization, whether in the spatial or spectral domain, generally enhances the quality of the generated outputs, with subtle differences in layout and texture. In contrast, omitting this step results in repetitive patterns and can even cause the model to fail as the generation nears completion. This issue is primarily due to error accumulation in the AR models, particularly in the raster-scan paradigm. Furthermore, increasing the gap between the original and

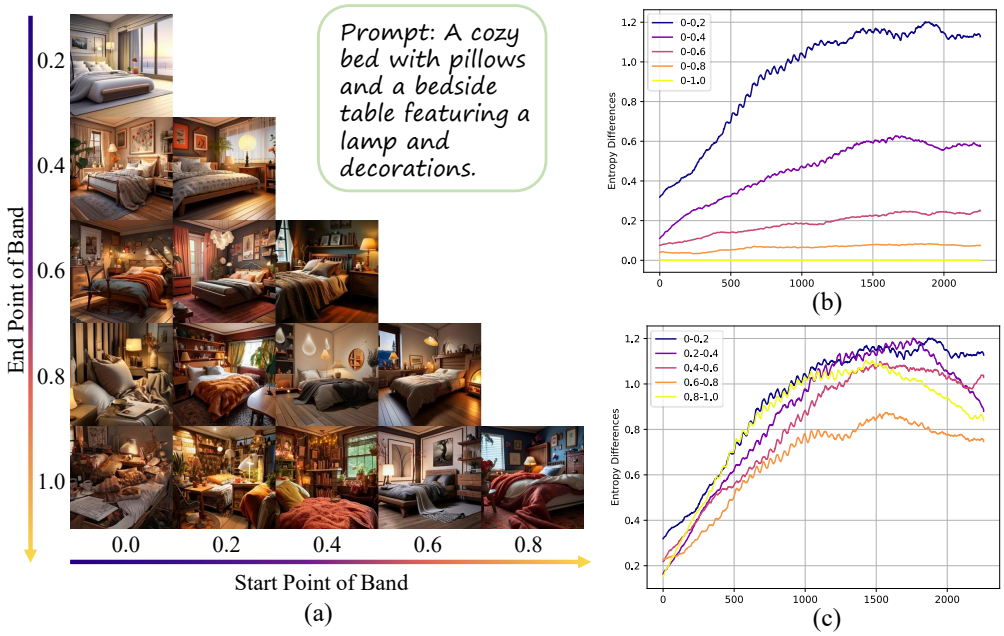


Figure 4: **Illustration of spectrum analysis.** (a) Visualizations under different spectral band selections. The horizontal axis represents the starting position of the spectral band, while the vertical axis represents the ending position. Bands located between them are retained for the weak model. Apparently, the 0-1.0 interval is equivalent to the baseline without any guidance. *Note that from the bottom left to the top right, the image becomes clearer as the bandwidth decreases.* (b) Entropy difference vs. different bandwidths with the same starting point. (c) Entropy difference vs. different start positions with the same bandwidth. The entropy difference is calculated by subtracting the baseline branch from the perturbation branch. Lumina-mGPT is used for demonstration.

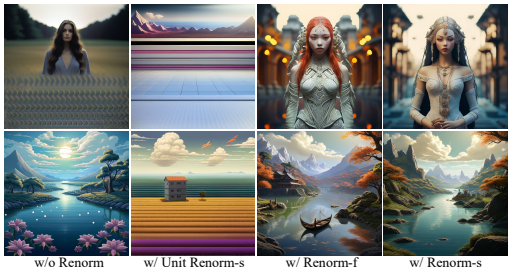


Figure 5: Lumina-mGPT performs unconditional generation under different renormalization strategies. Renorm-f and Renorm-s indicate the renormalization in spectral and spatial domain, as introduced in Eq. 10 and Eq. 11, respectively. Unit Renorm-s means use spatial normalization to unit length.

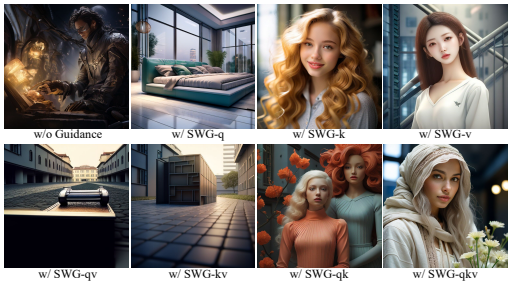


Figure 6: Lumina-mGPT performs unconditional generation employing various SWG strategies within the attention module. The term w/o guidance denotes the absence of any guidance mechanism. Configurations associated with SWG specify whether the query (Q), key (K), and value (V) representations are processed independently or jointly.

processed signals exacerbates this problem, resulting in visible artifacts, such as repetitive lines, as shown in the second column of the figure.

Ablation of Weak Branch Constructions. In addition to SWG, we design some other weak models and compare them in Tab. 4. Naively averaging the components or pruning specific layers has a limited effect on guiding. Specifically, these operations yield suboptimal results, with FID scores of 112.79 and 110.87, respectively. In contrast, SWG consistently outperforms on these metrics.

Table 4: Comparison of unconditional generation on COCO using NOVA, evaluated on 1K generated images. AS denotes aesthetic score. **Bold** indicates the best entries. Spatial and Spectrum indicate SWG with spatial and spectral renormalization, respectively.

	Baseline	Average	Pruning	Spatial	Spectrum
FID ↓	139.75	112.79	110.87	105.98	**104.51**
IS ↑	9.55	11.87	13.25	13.04	**13.44**
AS ↑	4.01	5.56	5.02	5.70	**5.72**



Figure 7: Applying SWG on different layers. Lumina-mGPT with a null prompt is used for demonstration. As the index of layers increases, the effectiveness of guidance decreases, and the synthesized results become blurry and distorted.

SWG demonstrates robustness to perturbations in the positional components. As illustrated in Fig. 6, the model fails for unconditional generation results without guidance, and the synthesized outputs appear blurry and distorted. In contrast, integrating SWG substantially enhances the visual quality of the generated results. SWG remains effective regardless of whether it is applied to the query, key, or value representations.

Interestingly, different perturbation locations can be combined flexibly. While these settings lead to changes in layout or contextual emphasis, the resulting output remains coherent and visually plausible. This consistency demonstrates the benefits of our spectrum-weakening design. The operation preserves the structural stability of intermediate representations, allowing the model to process them smoothly. It attenuates certain spectral components in the intermediate representations, yielding a weaker model variant that aligns with our view that the guidance effect does not rely on a particular structural configuration.

Layer Selection. Fig. 7 illustrates the impact of SWG application on the generated results at different layers. Intuitively, the higher the index number, the fewer layers it can affect, and the lower the guiding effect. Notably, applying SWG at the very beginning of the model seems to cause excessive differences between the weak model and the base model, resulting in a lack of detail in the final generated result, consistent with observations in Fig. 4.

5 Conclusion

This paper introduces SWG, a training-free visual guidance framework that builds a weak model by selectively retaining a subset of spectral components in intermediate representations, with a renormalization step to maintain feature scales and support stable generation. Demonstrating effective performance across various unconditional and conditional settings and AR modeling tasks, SWG extends the scope of CFG, offering a broader and more flexible paradigm for generation control.

Acknowledgment. This work was supported by Shenzhen Science and Technology Innovation Project (Grant No.2025110105) and NSFC (Grant No.42271426).

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, 2024.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[5] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021.

[8] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *NeurIPS*, 2022.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

[11] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021.

[16] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *NeurIPS*, 2024.

[17] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023.

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[19] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *CVPR*, 2025.

[20] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2024.

[21] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024.

[22] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[23] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

[24] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024.

[25] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *ACL*, 2023.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[28] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

[29] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[30] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *ICLR*, 2024.

[31] Charlie Nash, Joao Carreira, Jacob Walker, Iain Barr, Andrew Jaegle, Mateusz Malinowski, and Peter Battaglia. Transframer: Arbitrary frame prediction with generative models. *arXiv preprint arXiv:2203.09494*, 2022.

[32] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. In *CVPR*, 2025.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[35] Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. *arXiv preprint arXiv:2506.10036*, 2025.

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[38] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*, 2024.

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

[40] Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. *LAION. AI*, 2022.

[41] Youngwoo Shin, Jiwan Hur, and Junmo Kim. Ssg: Scaled spatial guidance for multi-scale visual autoregressive generation. In *ICLR*, 2026.

[42] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024.

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[45] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[46] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024.

[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv: 2307.09288*, 2023.

[48] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2023.

[49] Chaoyang Wang, Xiangtai Li, Henghui Ding, Lu Qi, Jiangning Zhang, Yunhai Tong, Chen Change Loy, and Shuicheng Yan. Explore in-context segmentation via latent diffusion models. In *AAAI*, 2025.

[50] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024.

[51] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[52] Hu Yu, Hao Luo, Hangjie Yuan, Yu Rong, and Feng Zhao. Frequency autoregressive image generation with continuous tokens. *arXiv preprint arXiv:2503.05305*, 2025.

[53] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[54] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023.

Overview. In the supplementary materials, we present more results and details:

- **A.** More details on the method.
- **B.** More empirical studies and visual results.
- **C.** Discussion of the broader impacts, safeguards and limitations.
- **D.** Compute Reporting.

A Method Details

Implementation Details. We apply spectrum selection to the attention module. By default, we retain the components located in the 0.25-0.5 spectral band. The guidance strength is set to 2 for Lumina-mGPT, 3 for NOVA, and 4 for Infinity, respectively. We use Qwen2.5-VL to generate captions for COCO. Unless otherwise specified, we evaluate on 30K generated images for COCO and compute FID on 10K generated images for ImageNet.

Reason for Renormalization. Recall that SWG's core operation is $\mathbf{y} = \mathbf{W}^{-1}\mathbf{M}\mathbf{W}\mathbf{x}$, with $\mathbf{W}$ and $\mathbf{M}$ defined in Eqs. 3 and 6, respectively. According to Parseval's theorem, we obtain $\|\mathbf{y}\|_2 = \|\mathbf{M}\mathbf{W}\mathbf{x}\|_2 \leq \|\mathbf{W}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$, with equality when the discarded spectral components have zero energy.

B Empirical Results

B.1 More Results on Class-condition Models

In this subsection, we provide more results on class-condition models, including random-order and scale-wise prediction.

For RandAR evaluation with 50K images, we perturb layers 8 and 9, using CFG and SWG scales of 4 and 0.2, respectively. As shown in Tab. 5, SWG is compatible with CFG and can further enhance the performance. Specifically, our SWG further improves the FID to 2.23. More visual results are in Fig. 8.

For VAR, we apply SWG on the layer 0-3 on VAR-d16, and set the guidance scale 2.0 for unconditional generation and 5.5 for conditional generation, respectively. As shown in Tab. 6, our proposed SWG consistently improves generation quality on scale-wise AR models.

Table 5: Performance of RandAR on class-conditional ImageNet 256×256 benchmark. 50K images are generated for evaluation, with a sampling step of 88.

	FID ↓	IS ↑	Precision ↑	Recall ↑
RandAR w/ CFG	2.25	317.77	0.80	0.60
+SWG	2.23	303.96	0.81	0.59

Table 6: Performance of VAR-d16 on class-conditional ImageNet 256×256 benchmark. 10K images are generated for evaluation on ImageNet.

	Guide	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Uncond	-	54.43	19.19	0.53	0.47
	SWG	41.96	23.34	0.58	0.45
Cond	-	14.49	104.15	0.71	0.66
	SWG	8.74	188.24	0.82	0.58

B.2 More Results on Text-condition Models

In this subsection, we provide more results on text-to-image models. Fig. 9 visualizes the effectiveness of our proposed SWG on Infinity. The synthesized results exhibit better detail and realism. Moreover, SWG can effectively improve the DPG-Bench score, as shown in Tab. 8.



Figure 8: More visual results of RandAR on ImageNet dataset.



(a) Unconditional Generation (b) Conditional Generation

Figure 9: More visual results of Infinity on COCO datasets. (a) Unconditional generation with null prompts. (b) Conditional generation with specific prompts. The prompts from left to right: "*A wooden bunk bed with a desk and drawers underneath*", "*Dog peeking out of a pontoon boat*", "*Bright green apples in a bowl*".

B.3 Inference Time Analysis

We analyzed the inference time. Due to DFT, our proposed SWG is slightly slower than CFG. However, since this operation is applied only to specific layers, the additional time cost is limited, as shown in Tab. 7.

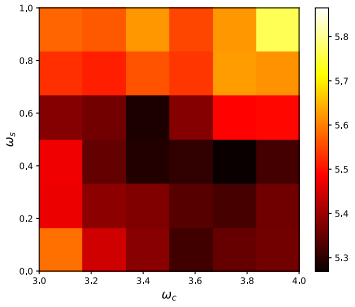


Figure 10: Compatibility between SWG (ω_s) and CFG (ω_c) on the ImageNet dataset. The heatmap shows FID scores, with darker colors indicating better performance.

Table 9: Compatibility of SWG and CFG on COCO using NOVA, evaluated on 1K generated images. CS and AS indicate CLIP score and aesthetic score, respectively. **Bold** indicates the best entries.

Method	FID ↓	IS ↑	CS ↑	AS ↑
NOVA w/CFG	67.59	25.67	**32.86**	5.80
+SWG	**67.02**	**26.30**	32.51	**5.96**

Table 7: Inference time comparison between SWG and CFG. RandAR 0.7B is used for evaluation with a batch size of 64.

	CFG	SWG
images / sec.	12.59	11.62

Table 8: Evaluation of SWG on Infinity using DPG-Bench. We report the overall score.

	-	SWG
Overall	73.43	77.03

Compatibility with CFG. We demonstrate both quantitatively and visually that SWG is compatible with CFG. As shown in Tab. 9 and Tab. 10, SWG is compatible with CFG for enhanced aesthetic score and overall GenEval performance. Fig. 10 further supports this finding, where the optimal configuration attains an FID of 5.26 at $\omega_c = 3.8, \omega_s = 0.4$, surpassing either component used alone.

Interestingly, while both SWG and CFG provide effective guidance, they emphasize different aspects. In Tab. 11, compared with the baseline method, both CFG and our SWG can consistently improve the generation quality. However, CFG yields the highest CLIP score, reflecting more substantial prompt alignment, whereas SWG attains higher aesthetic scores, placing greater emphasis on highlighting image sharpness and overall quality. Note that our proposed SWG is training-free, while CFG relies on unconditional predictions learned during training as the reference for guidance. Qualitative examples in Fig. 11 also show that the images synthesized by SWG have richer details and are more realistic.

B.4 Extension to Diffusion Models

SWG also extends to diffusion models. We evaluate it on SDXL by applying spectrum selection to query representations in the middle layers and retaining the last 10% of spectral components. Fig. 12 shows improved visual quality with an empty prompt, while Tab. 12 reports competitive performance against Smoothed Energy Guidance (SEG) in both unconditional and conditional generation.

B.5 Failure Cases

Our proposed method occasionally produces artifacts or fails to fully adhere to the instructions, as illustrated in Fig. 13.

C Discussion

C.1 Broader Impacts

This paper proposes a training-free and condition-free guidance approach for visual AR generation that improves image quality without retraining the underlying models. The method may benefit visual content creation, but enhanced image realism may also facilitate the creation of misleading content. As SWG operates on pretrained models, its outputs may inherit biases and limitations from the underlying training data.

C.2 Safeguards

Our method is training-free, which means it does not introduce new knowledge into the pre-trained models. Since training data heavily influences the generated content, ensuring its fairness and cleanliness is crucial to preventing harmful content.

C.3 Limitations

Despite the promising results of our method, it still has limitations. SWG sometimes fails to capture fine details or has artifacts in the generated results, suggesting room for improvement.

Table 10: Compatibility of SWG and CFG on GenEval.

Method	Single Obj	Two Obj	Counting	Colors	Pos	Color Attr	Overall
Infinity w/CFG	1.00	0.76	0.62	0.84	0.25	0.58	0.67
+SWG	1.00	0.76	0.60	0.86	0.27	0.62	0.68



Figure 11: Visualization of different guidance strategies on Lumina-mGPT. Compared to CFG, SWG is superior in overall image clarity and aesthetic quality.

Table 11: Comparison with CFG on COCO using Infinity, evaluated on 1K generated images. CS and AS indicate CLIP score and aesthetic score, respectively. Uncond. Train indicates whether guidance requires unconditional training support.

Uncond. Train	Guide	FID ↓	IS ↑	CS ↑	AS ↑
-	-	79.28	22.84	31.45	6.24
✓	CFG	77.17	25.22	32.29	6.25
✗	SWG	74.69	24.06	31.73	6.39

D Compute Reporting

All experiments are conducted on a single-node cluster equipped with Intel(R) Xeon(R) Gold 6248 CPUs @ 2.50GHz, 8 NVIDIA A100 GPUs, and 1.5 TB of system memory. Our method requires no training.

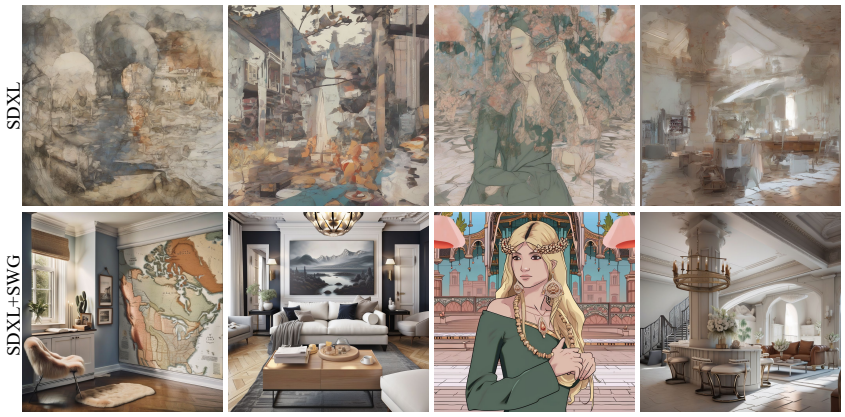


Figure 12: Visual results of applying SWG to SDXL with an empty prompt.

Table 12: Comparison with diffusion baselines on COCO dataset. SDXL is used for demonstration.

	Guide	FID ↓	IS ↑	CS ↑	AS ↑
Uncond	-	185.76	7.23	-	4.82
	SEG	138.14	9.90	-	6.09
	SWG	137.75	10.67	-	6.14
Cond	-	108.03	14.58	27.60	5.20
	SEG	75.02	20.06	30.12	6.14
	SWG	68.94	23.89	32.72	6.13



A display of green and yellow vegetables at a market.



A skateboarder performs a trick on a ramp.



A pizza with pepperoni and cheese served alongside a bowl of clams on a wooden table.

Figure 13: Failure case analysis of Lumina-mGPT. The prompts corresponding to each generated result are provided below.

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 main claims made in the abstract and introduction accurately reflect the paper's contributions and scope

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: in the appendix

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: our work focuses on method design and empirical evaluation, with supporting analysis in method section

Guidelines:

- The answer [N/A] 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: we provide implementation details

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: project resources are available on our GitHub page

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: we provide implementation details

   Guidelines:

   - The answer [N/A] 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 evaluation focuses on existing benchmark comparisons without error bars

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: we provide them in the appendix

   Guidelines:

   - The answer [N/A] 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 conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] 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: in the appendix

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: in the appendix

    Guidelines:

    - The answer [N/A] 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: creators are properly credited

    Guidelines:

    - The answer [N/A] 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 release new assets

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.