EAR: EFFICIENT AUTOREGRESSIVE TEXT-TO-IMAGE SYNTHESIS FROM A GAME-THEORETIC VIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional autoregressive text-to-image synthesis approach rely on next-token prediction, which would hinder inference efficiency due to their sequential, predefined token generation order. To enable high parallelization without sacrificing generation quality, we revisit autoregressive text-to-image models from a theoretical perspective by analyzing the generation order of visual tokens. We also propose the Efficient AutoRegressive text-to-image synthesis model (EAR), which provides a principled explanation of how visual tokens are generated in autoregressive models, while still supporting fine-grained image synthesis with efficient inference. Specifically, drawing inspiration from Game theory, we suggest employing Shapley value to evaluate the interactions between regions that arise from different order steps. Moreover, we present a novel text-to-image synthesis framework that facilitates rapid generation, with the guidance of both global representation and intermediate local embedding extracted from a pretrained text encoder. Extensive experiments, including quantitative evaluations and qualitative visual comparisons, on the public benchmarks show that our EAR model achieves lower latency than previous autoregressive text-to-image models without sacrificing image quality. We hope that the EAR model will offer novel insights into the interpretability of autoregressive models, thereby providing valuable benefits to researchers in this field.

1 INTRODUCTION

Autoregressive (AR) models have demonstrated outstanding performance and remarkable scalability Henighan et al. (2020) in large language models Achiam et al. (2023); Guo et al. (2025). Exemplified by the successful extension of "next-token prediction" to the visual domain, VQGAN Esser et al. (2021) and other models Sun et al. (2024); Xie et al. (2025); Zhou et al. (2025) have enabled significant advancements in autoregressive image generation. However, this design poses distinct challenges: while text inherently possesses causal dependencies, images are organized within a two-dimensional spatial domain, presenting AR models with a significant issue. Moreover, the methodology of generating a single token per step results in a memory-constrained workload, leading to latency that increases in proportion to the number of steps. Some researchers, e.g., Tian et al. (2024) in VAR, propose "next-scale prediction" via coarse-to-fine multi-scale representations to reduce generation steps and thus latency. Unfortunately, its multi-scale token representation exhibits a fundamental divergence from the commonly adopted flat token representation, rendering it incompatible with the widely employed flat vision perception foundation models (e.g., CLIP Radford et al. (2021) and DINO Oquab et al. (2023)), and thereby making it hard to work well with perception backbones, which have been demonstrated to be of paramount importance for the development of unified multimodal systems Jiao et al. (2025); Zhao et al. (2025); Lin et al. (2025); Wu et al. (2025).

To address these challenges, recent works, such as RandAR Pang et al. (2025), have endeavored to expedite the process by transitioning from next-token prediction to multi-patch prediction, but significant memory and computational costs are incurred due to the increased sequence length. Non-autoregressive models like MaskGIT Chang et al. (2022) has embraced a masked modeling strategy Devlin et al. (2019) to facilitate parallel token generation in a random sequence. However, the dependence on bidirectional attention prevents the utilization of the KV cache, making this approach less efficient than autoregressive methods. Infinity Han et al. (2025) performs text-to-image generation by predicting residuals of next scale which are conditioned on previous predictions and the text

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

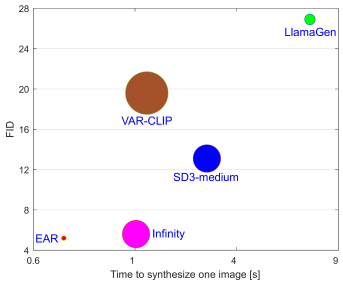

Figure 1: **Image synthesis quality and speed of different autoregressive models.** Each bubble's area is linearly proportional to the number of parameters of a model. The y-axis correlates with FID (at 512 × 512 resolution) evaluated on MJHQ 30K.

input in an autoregressive fashion. Nevertheless, Infinity would encounter the similar drawbacks as vanilla VAR Tian et al. (2024).

To overcome the limitations of previous models, we introduce EAR, a novel efficient auto-regressive text-to-image synthesis model that theoretically investigates visual token generation orders through Game theory by utilizing Shapley value, and thus enables flexible parallel autoregressive modeling. Based on this new framework, we further conduct text-to-image (T2I) synthesis by predicting discrete feature maps in a parallel manner, which is guided by both global representation and intermediate local embedding extracted from a pretrained text encoder. Consequently, our proposed EAR offers several advantages, as illustrated in Fig. 1. In particular, with the smallest number of parameters, EAR achieves at least 1.4 × lower latency compared to previous autoregressive models (e.g., Infinity Han et al. (2025), LIamaGen Sun et al. (2024), Parti-3B Yu et al. (2022)) without compromising image quality.

Overall, our major contributions are summarized as following: (1) To the best of our knowledge, we are the first to investigate how to formulate the schedule for the generation order of autoregressive models through the Game-theoretical analysis. We also devise an efficient method to accelerate the calculation of Shapley value based on mathematical induction. (2) We propose an efficient autoregressive text-to-image synthesis model, which can predict discrete feature maps in parallel, thereby significantly enhancing the generation speed. (3) Extensive evaluations reveal that our model offers competitive generation quality and outperforms others in throughput and memory use, creating a new benchmark for high-performance, high-efficiency autoregressive text-to-image synthesis.

2 Related Work

Text-to-image Synthesis. Text-to-image (T2I) models can be categorized into three types based on their probability distribution acquisition techniques: Generative Adversarial Networks (GANs) Zhao et al. (2023); Ye et al. (2023); Tan et al. (2023), Diffusion Model (DM) Dhariwal & Nichol (2021); Rombach et al. (2022), and AutoRegressive Model (AR) Hoffmann et al. (2022); Touvron et al. (2023). GANs often suffer from training instability and model collapse. Typical diffusion models such as SD3.0 Esser et al. (2024), SORA Brooks et al. (2024) and DALL-E 3 Betker et al. (2023) have shown unprecedented results in T2I tasks. However, enhancement of such models still faces hurdles, mainly because multi-stage denoising requires significant computational resources. AutoRegressive models synthesize images by predicting the next token from a codebook, with integration of Large Language Models (LLMs) like GPT Floridi & Chiriatti (2020), BERT Devlin et al. (2019), and LLaMA Touvron et al. (2023). The Vector Quantized Variational Autoencoder (VQVAE) Van Den Oord et al. (2017) or VQGAN Esser et al. (2021) are typically employed in interpreting image representation through discrete tokens. On the other hand, since discrete-value spaces are unnecessary for AutoRegressive models, Masked Autoregressive Models (MARs) Li et al. (2024b) have been proposed by leveraging the swiftness of sequence modeling. Specifically, MARs utilizes continuous token representations, employing a diffusion-based objective for optimization and a lightweight token-wise diffusion model for token sampling. Unlike next-token prediction and masked AR modeling, Tian et al. Tian et al. (2024) proposes a "next-scale prediction" approach named Visual Autoregressive Modeling (VAR) to surpass diffusion transformers in image generation quality. Infinity Han et al. (2025) extends the VAR to the conditioned text-to-image synthesis

paradigm. However, "next-scale prediction" strategy may be in conflict with well-known flat vision perception foundation models (e.g., CLIP Radford et al. (2021) and DINO Oquab et al. (2023)), limiting its broad applications. Thus, in this work, we re-examine the "next-token prediction" and propose an efficient autoregressive text-to-image synthesis model based on Game theory.

Parallelism in Sequence Image Generation. Parallel generation has been widely studied in the field of language modeling. Inspired by BERT Devlin et al. (2019) and based on encoder-only Yang et al. (2019) architecture with bidirectional attention, MaskGIT Li et al. (2023); Yu et al. (2023) facilitates the concurrent generation of multiple tokens in a single pass, employing a randomized ordering mechanism. Nevertheless, this masked sequence modeling lacks a key-value cache in bidirectional attention mechanisms. As a result, the efficiency of the inference process continues to be constrained. Another mainstream approach adopts a decoder-only transformers for causal modeling of image token sequences and autoregressive generation. For example, RandAR Pang et al. (2025) inserts positional instruction tokens to designate target positions, thereby achieving arbitrary generation orders. However, the positional instruction tokens double the memory consumption which would compromise the efficiency. ARPG Li et al. (2025) also employs an encoder-decoder framework, in which target-aware query tokens engage in cross-attention with the encoder's key-value cache. Unfortunately, since the target positions fail to contribute any key-value pairs, the image tokens generated in the same parallel step are produced independently. Hawk Chen et al. (2025) utilizes speculative decoding strategy introduced by Medusa Cai et al. (2024) and Eagle Li et al. (2024c). However, according to prior evaluations Xia et al. (2024), Medusa is not among the most efficient or state-of-the-art algorithms for text speculative decoding. More recently, LPD Zhang et al. (2025) introduces flexible parallelized autoregressive modeling and locality-aware generation ordering. However, LPD solely concentrates on simple class-conditional image generation, rather than engaging in more complex text-to-image synthesis. More importantly, the locality-aware generation ordering schedule is not evaluated by theoretical analysis and can not be guaranteed to be the optimal solution. Recent work PAR Wang et al. (2025) shows that independent sampling for strongly correlated local tokens can cause inconsistent generation and disrupted patterns. Therefore, the major objective of our paper is to theoretically formulate the generation ordering schedule, and then develop an efficient autoregressive text-to-image synthesis model.

3 Interpretability of Autoregressive Models

3.1 Preliminaries

Rethinking AutoRegressive Modeling. Decoder-only autoregressive models perform next-token prediction by splitting the image into patches and discretizing these patches via a tokenizer. Formally, given a 1D sequence of N tokens (denoted as $x = [x_1, x_2, \ldots, x_N]$) and input text T, the text-to-image autoregressive framework attempts to factorize the total joint distribution as:

$$p_\theta(x_1, x_2, \ldots, x_N; T) = \prod_{n=1}^{N} p_\theta(x_n | x_{<n}; T) \tag{1}$$

Consequently, the training target of the autoregressive model is to optimize the parametric approximations $p_\theta(x_n|x_{<n}; T)$ with respect to those one-step conditional distributions. However, during sampling, it causes N sequential steps and makes a big efficiency bottleneck.

To cut down on sequential generation steps, we can group tokens into G distinct sets $\{X_1, X_2, \ldots, X_G\}$, where each set $X_g = \{x_{g1}, x_{g2}, \ldots, x_{gm}\}$ is jointly predicted, thus yielding the following:

$$p_\theta(x_1, x_2, \ldots, x_N; T) = \prod_{g=1}^{G} p_\theta(X_g | X_{<g}; T) \tag{2}$$

Thus, the training objective turns to optimizing $p_\theta(X_g|X_{<g}; T)$. Earlier work reveals that grouping tokens in raster order directly causes notable performance drops Wang et al. (2025); Pang et al. (2025) since tokens adjacent in space have strong mutual dependencies, and independent sampling typically brings about generation inconsistencies inside a group. Therefore, it is necessary to break away from the raster order during token grouping. Additionally, the size of the prediction group

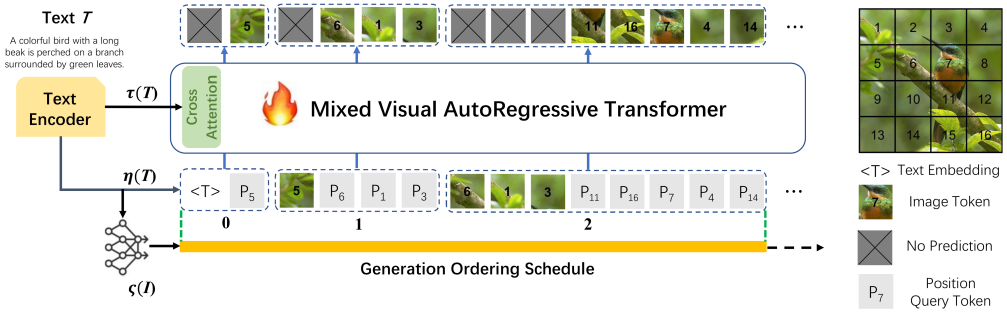

Figure 2: **The overview of our proposed EAR model.** Given an input text, EAR extracts a compact global representation and intermediate local embedding from a pretrained CLIP model, while simultaneously producing the generation ordering schedule through a neural network $\varsigma(I)$, which is guided by Game Theory proposed in Section 3. Using position query tokens and hybrid cross-attention, the mixed visual autoregressive transformer enables parallel generation of target tokens from a codebook produced by the pretrained VQVAE model.

X_g is required to increment progressively. With the growth of the context $X_{<g}$, it provides stronger conditioning, allowing more tokens to be predicted simultaneously. To address these challenges, we group the generation ordering schedule theoretically through Game-theory, and enable an arbitrary generation order and varying degrees of parallelization.

Shapley Value. In recent years, the Shapley value Shapley (1953) has been widely employed for feature attribution analysis in deep neural networks. Compared with other attribution methods Zhao et al. (2024), the Shapley value, grounded in cooperative Game theory, can provide a unique attribution value with a solid theoretical foundation for each input feature, thereby ensuring its fairness and credibility. Specifically, given a set of s input players, denoted as $S = \{1, 2, \cdots, s\}$, participating in a game v, these players can obtain a score $v(S)$. A game v maps each participating player to a numerical value, with the score when no players participate in the game denoted as $v(\emptyset)$. The total payoff obtained by s players can be expressed as $v(S) - v(\emptyset)$. When player i joins any potential subset $H \subseteq S \backslash \{i\}$, the Shapley value $\phi(i|S)$ is calculated as follows:

$$\phi_v(i|S) = \sum_{H \subseteq S \backslash \{i\}} \frac{|H|!(s-1-|H|)!}{s!} \left(v(H \cup \{i\}) - v(H)\right) \tag{3}$$

where the symbol '\' indicates exclusion. The key properties of the Shapley value is provided in Appendix A.1.

3.2 Problem Formulation

Let $\kappa = \{X_g \in R^D\}_{g=1}^G$ represents generation groups. For a special group X_{g_0}, from the attribution perspective of the Shapley value, the attribution diagrams $\Re(X_{g_0})$ are calculated as following:

$$\Re(X_{g_0}) = \phi_{v_{g_0}}(\cdot|\kappa) \tag{4}$$

where $\phi_{v_{g_0}}(\cdot|\kappa) \in R^G$ are computed by Shapley value. For example, $\phi_{v_{g_0}}(g_\gamma|\kappa) \in R$ is the Shapley value of the group g_γ. Since we utilize a codebook generated by the pretrained VQVAE Van Den Oord et al. (2017) model, the game v_{g_0} between g_0 and g_γ could be formulated through Mutual Information (MI) Kraskov et al. (2004) as:

$$\phi_{v_{g_0}}(g_\gamma|\kappa) = w_{0,\gamma} \cdot MI(g_0; g_\gamma) = w_{0,\gamma} \cdot \sum_{\alpha \in g_0} \sum_{\beta \in g_\gamma} P(\alpha, \beta) \log_2 \frac{P(\alpha, \beta)}{P(\alpha)P(\beta)} \tag{5}$$

where $w_{0,\gamma}$ is a distance-decay weighting across blocks, motivated by the observation that neighboring blocks typically exhibit higher mutual information than distant ones. A small weight can be assigned to neighboring blocks for balance. $w_{0,\gamma} = 1 - \exp(-\alpha_w \cdot d_{0,\gamma})$ and $d_{0,\gamma} = ||Ctr_0 - Ctr_\gamma||_2$. Here, Ctr_0 and Ctr_γ represent the spatial center coordinates of g_0-th group and g_γ-th group respectively. α_w is a hyperparameter, which is set to 0.1 in our experiments. The mutual information is estimated using a neural MI estimator based on the Donsker–Varadhan representation (i.e., MINE Belghazi et al. (2018)) for high-dimensional settings.

In other words, for any g_γ-th group, we expect the generated results to be contributed solely by the g_0-th group, provided that the g_0-th group emerges earlier than the g_γ-th group. Therefore, the Shapley value of the group g_γ is relatively higher. If the contrary is true (i.e., the g_γ-th group emerges earlier than the g_0-th group), it will have no impact, resulting in a lower Shapley value. Thus, we propose an interpretability autoregressive model loss function as:

$$L_{MI} = \mathrm{E}[\sum_{g_0} \sum_{g_\gamma \neq g_0} |\phi_{v_{g_0}}(g_\gamma|\kappa)|], \quad \text{if } g_0 \text{ is generated earlier than } g_\gamma \tag{6}$$

This loss function can be intuitively understood as follows: the g_0-th group only affects the g_γ-th group that satisfies the condition that the g_γ-th group emerges later than the g_0-th group, and has no influences on other groups that emerge earlier than the g_0-th group. Note that the relationship between the learned group generation schedule and the Shapley value is discussed in Appendix A.2.

3.3 Time Efficiency Optimization

Unfortunately, the calculation of the aforementioned loss function requires a time complexity of $O(G(G-1)2^G)$ in each training step of our proposed EAR Model. Specifically, a naïve implementation needs to compute the Shapley value of each group with respect to other groups. Such computational cost is extremely high, making this implementation impractical in real world applications. Therefore, based on the proposed Theorem 1, we introduce an equivalent and efficient method to simplify the calculation.

Theorem 1. If for any subset $H \subseteq S\backslash\{i\}$, the conditions $v(H \cup i) = v(i)$ and $v(H) = v(\emptyset)$ are satisfied, then all participants in $S\backslash\{i\}$ are considered dummy players.

Proof. Theorem 1 is proven by induction on the cardinality of set S.

Base Case: When $|S| = 2$ and $|H| = 1$, for $j \in H$, we have $v(\{j\} \cup \{i\}) = v(\{i\})$ and $v(\{j\}) = v(\{j\} \cup \{\emptyset\}) = v(\emptyset)$. This satisfies the dummy condition for participant j. Therefore, the only participant in H can be regarded as a dummy player.

Inductive Step: Assume that the hypothesis of Theorem 1 holds when $|S| = s \geq 2$. Then, when $S = s + 1$, randomly select a participant $j \in S\backslash\{i\}$, and treat $\{i, j\}$ as a single participant. In this way, for any subset $H \subseteq S\backslash\{i, j\}$, we obtain:

$$\begin{aligned} &v(H \cup \{i, j\}) = v(\{H \cup \{j\}\} \cup \{i\}) = v(\{i\})//\{H \cup \{j\}\} \subseteq S\backslash\{i\}, \\ &v(\{i, j\}) = v(\{j\} \cup \{i\}) = v(\{i\})//\{j\} \subseteq S\backslash\{i\}. \end{aligned} \tag{7}$$

Then, we have:

$$v(H \cup \{i, j\}) = v(\{i, j\}) \tag{8}$$

Furthermore, we obtain:

$$v(H) = v(\varphi)//H \subseteq S\backslash\{i, j\} \tag{9}$$

Therefore, we receive:

$$H \subseteq S\backslash\{i\} \tag{10}$$

In this way, we can conclude, based on the inductive hypothesis, that all participants in $S\backslash(i, j)$ are dummy players. Furthermore, since participant j is arbitrarily selected from $S\backslash\{i\}$, by considering all possible participants j, we can infer that all participants in $S\backslash\{i\}$ are dummy players. **The proof of Theorem 1 is completed**.

By constraining certain Shapley value to zero, the loss function in Equation 6 could essentially be regarded as enforcing the requirement that: When the g_γ-th group emerges later than the g_0-th group, for the set of groups $\upsilon = \{g_\gamma \in \kappa\}$, υ must serve as a dummy player for g_0. This goal can be achieved in a more efficient manner through Theorem 1, without requiring $G(G-1)$ computations on Shapley value. By using the following Equation, we could reduce the time complexity from $O(G(G-1)2^G)$ to $O(G2^G)$:

$$L_{MI} = \mathrm{E}\left[\sum_{g_0} \sum_{H \subseteq \kappa\backslash\upsilon} \left((v_{g_0}(H \cup \upsilon) - v_{g_0}(\upsilon))^2 + (v_{g_0}(H) - v_{g_0}(\emptyset))^2\right)\right] \tag{11}$$

According to Theorem 1, this loss function is expected to force each participant in the set $\kappa\backslash\upsilon$ to act as a dummy player, thereby achieving the design goal of the interpretability property for

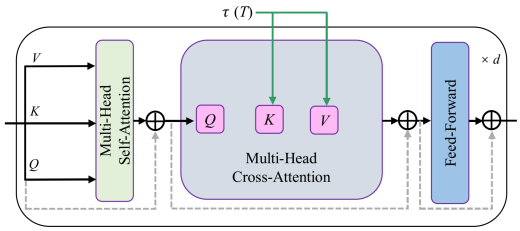

Figure 3: **The architecture of our proposed MVAT.** It consists of a multi-head self-attention, a novel multi-head cross-attention and a feed-forward network.



Figure 4: **Illustrative examples of EAR and its variant EAR-wot.** Images are synthesized at 512 × 512 resolution.

autoregressive models. However, the exact computation of L_{MI} still requires a time complexity of $O(G2^G)$ per step. To further address the issue of limited time resources, researchers adopt a polynomial sampling algorithm Castro et al. (2009) to accelerate the computation of Shapley value, reducing the computational complexity to $O(GK)$. Here, K denotes the number of samples, and would be evaluated in Appendix A.3.

4 EAR Model Details

Model Overview. As shown in Fig.2, Our text-to-image synthesis framework comprises three components: a pretrained text encoder such as Contrastive Language-Image Pre-Training (CLIP), a neural network producing the generation ordering schedule and an efficient autoregressive transformer (EAR). During the training procedure, we initially extract features from a frozen CLIP text encoder to generate conditioning vector $\tau(T)$ and $\eta(T)$. Meanwhile, we use a neural network (e.g., ResNet18) to produce the generation order for visual tokens, guided by the loss induced from Game Theory (denoted as L_{MI}). Ultimately, we utilize the proposed mixed visual autoregressive transformer (MVAT) to parallelly generate visual tokens. Note that more discussions about generation ordering schedule are provided in Appendix A.4.

Textual Extraction. In this work, for a given text prompt that reflects users' desires for image generation, our EAR model employs a pretrained text encoder such as Contrastive Language-Image Pre-Training (CLIP) Radford et al. (2021) to derive an intermediate local representation $\tau(T) \in R^{L\times d}$, where L represents the sequence length and d denotes the feature dimensionality, and a compact global representation $\eta(T) \in R^d$. While $\tau(T)$ enables fine-grained conditional guidance throughout the entire generation process, $\eta(T)$ captures the high-level structural objectives conveyed by the text prompt. As depicted in Fig.2, we use the $\eta(T)$ as a start token. Meanwhile, we extract the $\tau(T)$ through the identical CLIP text encoder to capture fine-grained textual details. $\eta(T)$ and $\tau(T)$ are fed into our mixed visual autoregressive transformer to enhance guidance of image generation (The technique details are provided in next section). In summary, this hybrid representation strategy allows the EAR model to generate highly accurate images conditioned on text, which exactly match the preferences specified by the user.

Mixed Visual Autoregressive Transformer. It is our goal to ensure a global understanding throughout the entire generation process while retain fine-grained details. Specifically, as shown in Fig. 3, we construct a cross-attention for feeding in intermediate local feature $\tau(T)$. Then, the output of cross attention corresponding to feature $\tau(T)$ can be formulated as:

$$Q = W_Q \cdot \Upsilon, \quad K = W_K \cdot \tau(T), \quad V = W_V \cdot \tau(T), \quad \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{12}$$

where Υ represents output vector derived from self-attention module. W_Q, W_K, W_V are learnable projection layers, and d is associated with the number of channels in Υ.

Loss Function. Given that we possess matched input-output data pairs (embedding→text), our goal is to maximize the likelihood of the corresponding image token as:

$$L_{ET} = \mathrm{E}_{x\sim p(x)}[-\log p(y|T)] \tag{13}$$

where $p(y|T)$ denotes the probability that the model predicts the correct token sequence given the input texts T, and x denotes the real image tokens. Then, the final loss function used for training the interpretable autoregressive models is defined as follows:

$$L = L_{ET} + \lambda_0 L_{MI} \tag{14}$$

where λ_0 is a hyperparameter and we fix it to 1 for training process, and L_{ET} denotes the standard direction-sensitive autoregressive training loss.

5 Experiments

5.1 Implementation Details

Models. We train EAR on text-to-image generation task using NVIDIA A100 GPUs, with approximately 20M image-text pairs from LAION and 10M internal images. The internal images have been recaptioned by Chen et al. (2024) to improve alignment between visual and textual data. To enable fair comparisons against existing auto-regressive image generation approaches, we adopt the LlamaGen tokenizer Sun et al. (2024), which features a codebook size of 16384 and a downsample factor of 16. We follow a standard decoder-only transformer architecture. Please refer to the Appendix A.5 for more details.

Training. By using both the LAION and internal image sets, the model is first trained at a resolution of 256 × 256 to ensure a diverse and comprehensive dataset for this resolution. We use AdamW Loshchilov & Hutter (2019) optimizer with (β_1, β_2) as (0.9, 0.95) and the weight decay of 0.05. A standard token dropout of 0.1 is applied. With regards to 512-resolution models, we first load pretrained 256-resolution models. Subsequently, we continue the training process on internal data, as it provides higher quality and greater consistency in the annotations needed for high-resolution image generation.

Evaluation Testing. To thoroughly assess the performance of our models, we employ well-established automated metrics like CLIP Score Hessel et al. (2021), ImageReward Xu et al. (2023), PickScore Kirstain et al. (2023), FID Heusel et al. (2017) and GenEval Ghosh et al. (2023). We first generate images at the native resolution of each evaluated model, and then resize them to 512 × 512. For more details about the evaluation process, please refer to Appendix A.6. We compute metrics using two widely adopted text-to-image datasets: MS-COCO Lin et al. (2014) and MJHQ Li et al. (2024a). For both datasets, we generate one image for every prompt in its 30,000-item validation prompt set.

5.2 Quantitative and Qualitative Evaluation

We compared EAR with several other state-of-the-art T2I models, including types of Diffusion Models and Autoregressive Models. From Table 1, we find that our proposed EAR achieves comparable results to the other models across all metrics with fewer model parameters. For example, our proposed EAR attains the lowest FID on both COCO and MJHQ datasets, demonstrating that our model can generate more realistic images in various scenarios. For other metrics like the PickScore, our proposed EAR continues to deliver results on par with those of other competing models. Note that all models reported in Table 1 synthesize images at a resolution of 512 × 512. We provide a human evaluation and additional quantitative results in Appendices A.7 and A.9, respectively.

With regards to Qualitative Evaluation, we contrast the synthesized images produced by Infinity Han et al. (2025), STAR Ma et al. (2024), SWITTI Voronov et al. (2025), SD3-medium Esser et al. (2024) and our EAR. We provide qualitative comparisons in Figures 5 and 9. As shown in Figure 5, our EAR is capable of synthesizing realistic images that in accordance with text instructions. It is strange that Infinity Han et al. (2025) could not produce better visualization results as depicted in the third and fourth rows of Figure 5. The reason we speculate is that infinity may be particularly sensitive to noisy, ambiguous, or insufficiently specified textual conditions. Such conditioning may leave the early coarse-scale predictions under-constrained, allowing structural errors introduced at early stages to propagate to subsequent scales.

Table 1: **Quantitative comparison among different models on COCO and MJHQ datasets.** Note that the results are taken from the most recent papers. * result is with prompt rewriting. Best results are in bold.

Model	Type	Parameters(B)	COCO 30K				MJHQ 30K				
			PickScore ↑	CLIP ↑	IR ↑	FID ↓	PickScore ↑	CLIP ↑	IR ↑	FID ↓	GenEval ↑
SDXL-Turbo Sauer et al. (2024)	DM	2.6	0.229	0.355	0.83	17.6	0.216	0.365	0.84	15.7	0.55
DMD2 Yin et al. (2024)	DM	2.6	**0.231**	0.356	0.87	14.3	0.219	0.374	0.87	7.2	0.58
SDXL Podell et al. (2024)	DM	2.6	0.226	0.360	0.77	14.4	0.217	0.384	0.78	7.6	0.55
SD3-medium Esser et al. (2024)	DM	2.0	0.227	0.354	1.01	19.5	0.215	0.363	0.91	13.1	0.65
Lumina-Next Zhuo et al. (2024)	DM	2.0	0.224	0.329	0.55	18.4	0.216	0.353	0.75	5.9	0.47
LlamaGen Sun et al. (2024)	AR	0.8	0.208	0.274	-0.25	44.8	0.194	0.288	-0.45	26.9	0.32
HART Tang et al. (2026)	AR	0.7	0.223	0.341	0.75	20.9	0.216	0.366	0.84	5.8	0.55
SWITTI Voronov et al. (2025)	AR	2.5	0.227	0.356	0.95	13.2	0.217	0.381	**0.93**	5.5	0.62
Infinity Han et al. (2025)	AR	2.0	0.230	**0.362**	**1.06**	13.9	0.220	0.381	0.91	5.6	**0.73** *
VAR-CLIP Zhang et al. (2024)	AR	3.1	0.215	0.279	0.34	18.3	0.199	0.287	0.66	19.6	0.49
STAR Ma et al. (2024)	AR	1.7	0.229	0.361	1.03	12.8	**0.221**	0.291	0.90	5.7	0.55
EAR	AR	0.5	0.229	**0.362**	1.05	**11.9**	0.219	**0.389**	0.91	**5.2**	0.62

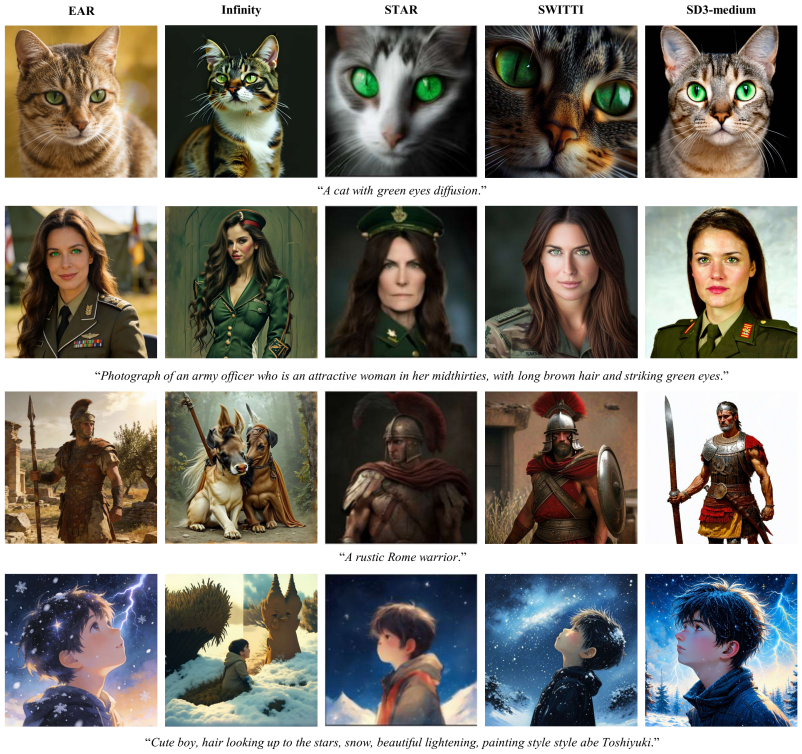

Figure 5: **Qualitative comparison of EAR for 512 × 512 generation against several state-of-the-art models.** More examples are in Appendix A.8.

5.3 Efficiency Profiling

We further analyze the efficiency of EAR's sampling and compare it to several models. All efficiency results are obtained on a single NVIDIA A100 GPU using BFloat16 precision. Similar to the evaluation strategy proposed in Wang et al. (2025), latency is measured with a batch size of 1, and throughput with a batch size of 16. We report the average latency over 500 inference steps after a 100 step warm-up.

As shown in Table 2, when considering the Latency metrics, EAR is among the most efficient image synthesis models, being more than 2 times faster than STAR Ma et al. (2024) and 5 times faster than SD3-medium Esser et al. (2024). SWITTI Voronov et al. (2025) has the lowest Latency, which is attributed to its adoption of a non-causal architecture. However, EAR attains the highest throughput, as it guarantees visibility across all concurrently predicted target positions and exclusively stores the

Table 2: **Comparison of 512 × 512 image generation times.** Best results are in bold.

Model	#Step	Latency(s) ↓	Throughput(img/s) ↑
SD3-medium Esser et al. (2024)	28	3.12	0.21
SWITTI Voronov et al. (2025)	14	**0.57**	15.72
Infinity Han et al. (2025)	10	1.03	13.89
VAR-CLIP Zhang et al. (2024)	10	1.35	12.79
STAR Ma et al. (2024)	10	1.51	11.62
EAR	48	0.72	**32.68**

generated tokens in the KV cache—thereby minimizing latency and maximizing throughput. Note that the recent PAR model Wang et al. (2025) is limited to generating class-conditional images on the ImageNet 256 $\times$ 256 benchmark, achieving a throughput of 6.83 images per second, which is still slower than our EAR operating at 512 $\times$ 512 resolution.

6 Ablation Study

We investigate the effect of our architectural choices on the image generation quality. In our paper, we propose a novel mixed visual autoregressive transformer (MVAT) in our EAR. Compared with other models (e.g., Infinity Han et al. (2025)) that fuse unique text-embeddings into autoregressive models, our EAR involves not only the compact global representation but also fine-grained textual guidances (See Equation (12)). Fig. 4 shows visual effects of our EAR with or without fine-grained textual guidances (denoted as EAR-wot). It seems that EAR could synthesize more reasonable visual details when compared with EAR-wot. For example, as shown in the first column of Fig. 4, the hair of the cat is denser and more realistic when produced by the original EAR model. In the second column of Fig. 4, the EAR-wot variant generates a cute boy while he stares straight ahead.

One of the key contributions of our paper is the use of Game Theory to monitor the generation ordering schedule through a neural network. Here, we evaluate the impact of our strategy. The corresponding quantitative results are listed in Table 3, where the term EAR-loc denotes an EAR variant that uses the locality-aware generation ordering strategy proposed in Zhang et al. (2025), EAR-rad refers to an EAR variant using random grouping, and EAR-ras designates an EAR variant that adopts raster grouping with progressively increasing group sizes. From Table 3, we can conclude that the introduction of Game Theory in our EAR can not only release a novel interpretability view to autoregressive models but also benefit the visual quality of produced images. Note that more visual results are given in Appendix A.10.

Table 3: **Quantitative comparisons among different generation orderings of EAR.** All tests are conducted on COCO and MJHQ datasets. Best results are in bold.

Model	COCO 30K				MJHQ 30K				
	PickScore ↑	CLIP ↑	IR ↑	FID ↓	PickScore ↑	CLIP ↑	IR ↑	FID ↓	GenEval ↑
EAR	**0.229**	**0.362**	**1.05**	**11.9**	**0.219**	**0.389**	**0.91**	**5.2**	**0.62**
EAR-loc	0.226	0.359	1.02	12.5	0.217	0.382	0.89	5.5	0.60
EAR-rad	0.202	0.330	0.88	15.7	0.196	0.351	0.72	6.0	0.51
EAR-ras	0.210	0.341	0.95	14.6	0.206	0.363	0.80	5.7	0.57

7 Conclusion

In this work, we introduce an efficient, interpretable autoregressive text-to-image synthesis model (EAR) from a Game-Theoretic perspective. Particularly, we utilize the Shapley value of Game Theory to evaluate the interactions between different regions, and then monitor the generation ordering schedule through a neural network. We further propose a novel text-to-image synthesis framework that facilitates rapid generation, with the guidance of both global and local features extracted from a pretrained text encoder. Extensive experiments substantiate the design decisions made for our model and demonstrate its remarkable image generation proficiency when compared with different types of state-of-the-art models.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we have not used generative AI tools at all.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work complies with the ICLR Code of Ethics. All experiments use publicly available benchmark datasets and environments and involve neither human participants nor personally identifiable information. No additional data were collected, and the study does not contain sensitive or private data. The contributions are methodological and are intended to involve Game Theory in machine learning. We have aimed to present the results accurately, discuss the limitations of the study, and adhere to established standards of research integrity. We declare no conflicts of interest or ethical issues beyond those ordinarily associated with responsible research practice.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We have taken extensive measures to facilitate the reproducibility of our results. The main paper and appendices provide detailed accounts of the proposed methods, underlying assumptions, and theoretical derivations. We also report the hyperparameters, model architectures, and training setups in full. The data preprocessing pipeline and evaluation procedures are clearly documented, and all datasets are publicly accessible. Collectively, these materials enable independent researchers to reproduce and validate our findings.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, 2024.

Zhi-Kai Chen, Jun-Peng Jiang, Han-Jia Ye, and De-Chuan Zhan. Hawk: Leveraging spatial context for faster autoregressive text-to-image generation. *Advances in Neural Information Processing Systems*, 38, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Pradeep Dubey and Robert James Weber. Probabilistic values for games. Technical report, 1977.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

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 *Forty-first International Conference on Machine Learning*, 2024.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *Advances in neural information processing systems*, 35:30016–30030, 2022.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3600–3610, 2025.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Haopeng Li, Jinyue Yang, Guoqi Li, and Huan Wang. Autoregressive image generation with randomized parallel decoding. *arXiv preprint arXiv:2503.10568*, 2025.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024b.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. 2024c.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Biye Li, Huaian Chen, and Yi Jin. Star: Scale-wise text-conditioned autoregressive image generation. *arXiv preprint arXiv:2406.10797*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 45–55, 2025.

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 *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

L. S. Shapley. A value for n-person games. *annals of mathematical studies*, 1953.

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.

Hongchen Tan, Baocai Yin, Kun Wei, Xiuping Liu, and Xin Li. Alr-gan: Adaptive layout refinement for text-to-image synthesis. *IEEE Transactions on Multimedia*, 25:8620–8631, 2023.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. In *International Conference on Learning Representations*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

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 preprint arXiv:2307.09288*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12955–12965, 2025.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. *arXiv preprint arXiv:2401.07851*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Senmao Ye, Huan Wang, Mingkui Tan, and Fei Liu. Recurrent affine transformation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 26:462–473, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, and Burcu Karagol Ayan. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024.

Zhuoyang Zhang, Luke J Huang, Chengyue Wu, Shang Yang, Kelly Peng, Yao Lu, and Song Han. Locality-aware parallel decoding for efficient autoregressive image generation. *arXiv preprint arXiv:2507.01957*, 2025.

Liang Zhao, Pingda Huang, Tengtuo Chen, Chunjiang Fu, Qinghao Hu, and Yangqianhui Zhang. Multi-sentence complementarily generation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 2023.

Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4): 99, 2024.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.

A APPENDIX

A.1 KEY PROPERTIES OF THE SHAPLEY VALUE

It is noteworthy that the Shapley value is the unique unbiased method for calculating contribution that satisfies four key properties Dubey & Weber (1977):

Dummy Player Property. If for any subset $H \subseteq S \backslash \{i\}$, the condition $v(H \cup \{i\}) - v(H) = 0$ holds, then participant i is considered a dummy player. The formula for calculating their contribution value is $\phi_v(i|S) = v(\{i\}) - v(\emptyset)$. This property implies that participant i makes no substantive contribution to the game.

Linearity Property. Suppose there exist three games u, v, and w, where game w is formed by combining games u and v, and satisfies $w(H) = u(H) + v(H)$, then the Shapley value of participant i in game w can be obtained by summing their Shapley values in games u and v, that is, $\phi_w(i|S) = \phi_u(i|S) + \phi_v(i|S)$.

Symmetry Property. If participants i and j make identical contributions to the game, then their Shapley values in game v should be equal. Specifically, if for any subset $H \subseteq S \backslash \{i, j\}$ the condition $v(H \cup \{i\}) - v(H \cup \{j\}) = 0$ holds, then we have $\phi_v(i|S) = \phi_v(j|S)$.

Efficiency Property. The sum of the Shapley value of all participants is equal to the total payoff obtained by all participants, that is $\sum_i \phi_v(i|S) = v(S) - v(\emptyset)$.

A.2 ALIGNMENT BETWEEN LEARNED GENERATION SCHEDULE AND SHAPLEY VALUE

To validate the game-theoretic motivation, we analyze the learned group generation schedule and compare it with the Shapley value based ranking. For each group, we compute its Shapley value by using Equation (5). The Shapley values based ordering is obtained by sorting groups in descending order of their Shapley values.

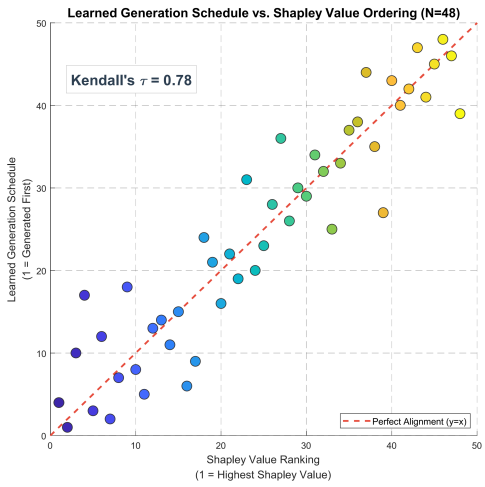

Figure 6: **Correlation analysis between the learned generation schedule and the Shapley value based ordering.** We visualize the rank alignment for N = 48 groups (512 $\times$ 512 resolution). The scatter plot exhibits a strong positive correlation with a Kendall's rank correlation coefficient of τ = 0.78. The data points clustering near the diagonal (dashed red line) empirically demonstrate that the model learns to prioritize groups with higher Shapley values (rank 1), generating them earlier in the schedule.

We then measure the rank correlation between the learned schedule and the Shapley-based ordering using Kendall's rank correlation coefficient τ Kendall (1938). As shown in Fig. 6, the learned schedule exhibits a strong positive correlation with the Shapley ranking (average τ = 0.78), indicating that groups with higher Shapley values tend to be generated earlier. This alignment empirically

supports that the learned ordering is consistent with the Shapley values based intuition rather than being arbitrary.

A.3 Sampling Strategy for Shapley Value Computing Acceleration

As discussed in Section 3.4, we utilize a polynomial sampling algorithm Castro et al. (2009) to accelerate the computation of Shapley values, reducing the computational complexity from $O(G(G-1)2^G)$ to $O(GK)$. For instance, in order to synthesize 512 × 512 resolution images, we fix the generation steps to 48, and thus G = 48. Consequently, the computational complexity is largely reduced from $O(48 \times 47 \times 2^{48})$ to $O(48 \times K)$, where we empirically set $K = 10^6$.

In this section, we further investigate how the different values of K affect the performance of our EAR. Specifically, we vary the value of K (from $K = 10^6$ to $K = 10^7$) and test its corresponding FID. The experimental results are shown in Figure 7. It is obvious that using a larger sampling size only results in a marginal FID improvement, while at the cost of more time-consuming computation. For example, the latency of $K = 10^7$ is 11 times higher than that of $K = 10^6$. Thus, we fix $K = 10^6$ in our EAR.

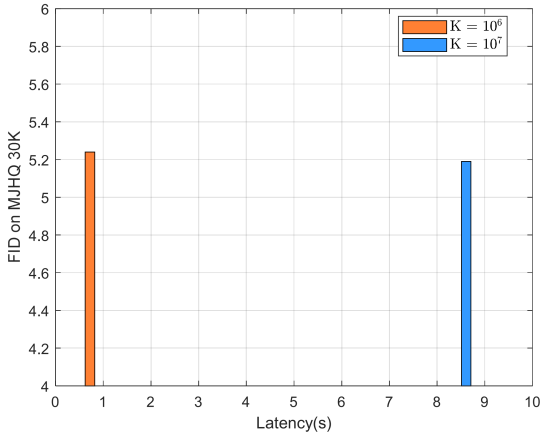

Figure 7: **Analyzing the influence of different K values on performance.** We use FID and Latency metrics for comparison.

A.4 Details of the Generation Ordering Schedule

In this paper, we use a neural network $\varsigma(I)$ to predict generation ordering schedule. Considering the different resolutions of the synthesized image, we fix the total generation steps as follows:

For 256 × 256 image resolution, the 20 generation steps are fixed as: [1, 2, 4, 5, 7, 8, 10, 11, 12, 14, 15, 16, 17, 18, 18, 19, 19, 20, 20, 20]

For 512 × 512 image resolution, the 48 generation steps are fixed as: [1, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 23, 24, 25, 26, 26, 27, 28, 28, 29, 29, 30, 30, 31, 31, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 34]

Thus, $\varsigma(I)$ functions as a classification network, dedicated to determining the category of each visual token (with 20 or 48 categories available). Note that, as mentioned before, we first train a model for 256 × 256 resolution, and then continue training for the 512 × 512 resolution model.

However, we can select different generation steps to produce images. The following paragraphs give an in-depth analysis of the number of generation steps in our model. Taking the 512 × 512 image resolution as an example, we may have:

512-a, 32 generation steps:

[1, 4, 6, 9, 11, 13, 16, 18, 20, 23, 25, 27, 29, 31, 33, 35, 36, 38, 40, 41, 42, 44, 45, 46, 47, 48, 48, 49, 49, 50, 50, 50]

512-b, 40 generation steps:

[1, 2, 4, 6, 7, 9, 10, 12, 13, 15, 16, 18, 19, 20, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 35, 36, 37, 37, 38, 38, 39, 39, 39, 40, 40, 40, 40]

512-c, 48 generation steps:

[1, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 23, 24, 25, 26, 26, 27, 28, 28, 29, 29, 30, 30, 31, 31, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33, 34]

512-d, 64 generation steps:

[1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 6, 7, 8, 8, 9, 9, 10, 10, 11, 12, 12, 13, 13, 14, 14, 15, 15, 16, 16, 17, 17, 18, 18, 18, 19, 19, 20, 20, 20, 21, 21, 21, 22, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25]

512-e, 92 generation steps:

[1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13, 14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 17, 17, 17, 17, 17, 17]

We further investigate the role of the number of generation steps in our model. To this end, we arrange experiments on the MJHQ 30K dataset and measure their FID values. From Figure 8, it has an obvious trend that at first the FID decreases sharply when using more generation steps. However, it becomes almost flat when taking additional generation steps. Therefore, in order to save computing time and keep high image quality, we fix 48 generation steps for 512×512 resolution in this paper.

It is worth noting that we conduct similar experiments on 256×256 resolution, and fix 20 generation steps for this case.

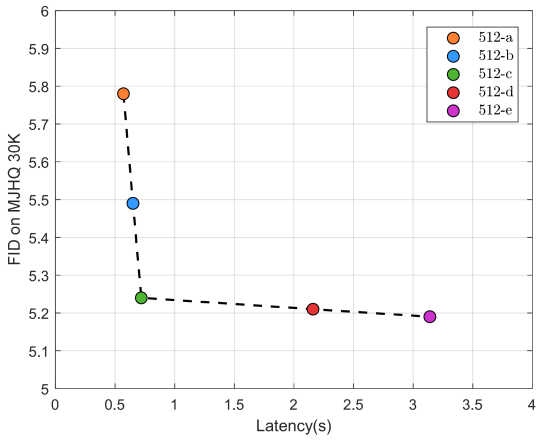

Figure 8: **Analyzing the influence of different generation steps on performance.** We use FID and Latency metrics for comparison.

A.5 Model Architecture

The configuration details of the model architecture are presented in Table 4. our model is built on a standard decoder-only transformer framework, and it has three key parameters: the number of layers, the hidden size, and the number of attention heads. Then, we could modify these three key parameters to construct three types of models, as shown in Table 4.

We additionally conduct experiments on COCO 30K dataset to compare the performance. The corresponding results are listed in Table 5. From Table 5, we observe that the larger-scale variant such as EAR-XL may marginally benefit the image quality of synthesized images but results in lower Throughput. Thus, considering the tradeoff between image quality and efficiency, we select the EAR

Table 4: **Model architecture configurations.** We alter the number of layers, hidden size and attention heads to construct three variants.

Model	Parameters	Layers	Hidden Size	Heads
EAR	445M	12	1024	12
EAR-L	839M	36	1280	20
EAR-XL	1.5B	48	1536	48

variant in this paper. Note that Table 5 presents a diagnostic ablation study. It should be interpreted only as a comparison under fixed data and a fixed training recipe, designed to examine the effect of increasing model capacity under controlled training conditions, rather than as a definitive scaling study.

Table 5: **Quantitative comparison among different models on COCO 30K datasets.**

Model	Latency(s) ↓	Throughput(img/s) ↑	COCO 30K			
			PickScore ↑	CLIP ↑	IR ↑	FID ↓
EAR	0.72	32.68	0.229	0.362	1.05	11.9
EAR-L	1.05	17.63	0.229	0.362	1.05	11.9
EAR-XL	1.41	10.58	0.229	0.363	1.05	11.8

A.6 Evaluation Details

To evaluate the alignment between images and text, we compute the CLIP Score Hessel et al. (2021) by leveraging features derived from the pretrained CLIP-ViT-H-14-laion2B-s32B-b79K encoder Schuhmann et al. (2022). We calculate FID on 30,000 generated images through downsampling stage: reduce to 256×256 using Lanczos interpolation (aligning with FID computation practices on COCO2014). Additionally, real data statistics are derived from all validation set images—around 40,000 from COCO2014 and 30,000 from MJHQ. With regards to GenEval, we synthesize 4 images per evaluation prompt (totaling 533 prompts). Thereafter, we adhere to the original evaluation protocol and utilize Mask2Former Cheng et al. (2022) with a Swin-S backbone as the object detector.

A.7 Human Evaluation

Despite the widespread adoption of automated metrics for text-to-image model evaluation, we argue that they lack the ability to fully represent all facets of image generation performance. Following the setup of the prior DreamBooth model Ruiz et al. (2023), we designed a user study to assess two key metrics: subject fidelity and prompt fidelity. We distributed questionnaires—each containing 20 comparative questions—to 60 participants, with 3 users completing the same questionnaire. This setup yielded a total of 1200 responses, covering the evaluation of 400 image pairs.

For subject fidelity, we randomly selected samples from a large dataset. Each question presented a group of real images of a specific subject, paired with one AI-generated image of the same subject (produced by each model using a randomly selected prompt). Participants were asked to answer: "Which of the two images most accurately reproduces the identity of the reference subject, including its category and specific details?"

As for prompt fidelity, we adopted a similar approach. Here, participants were prompted with: "Which of the two images aligns most closely with the description given in the reference text?"

We employed majority voting to aggregate the results, and the averaged outcomes are presented in Table 6. As indicated in this table, our data demonstrate a clear preference for EAR across both subject fidelity and prompt fidelity—consistent with the visual observations provided in Fig. 9.

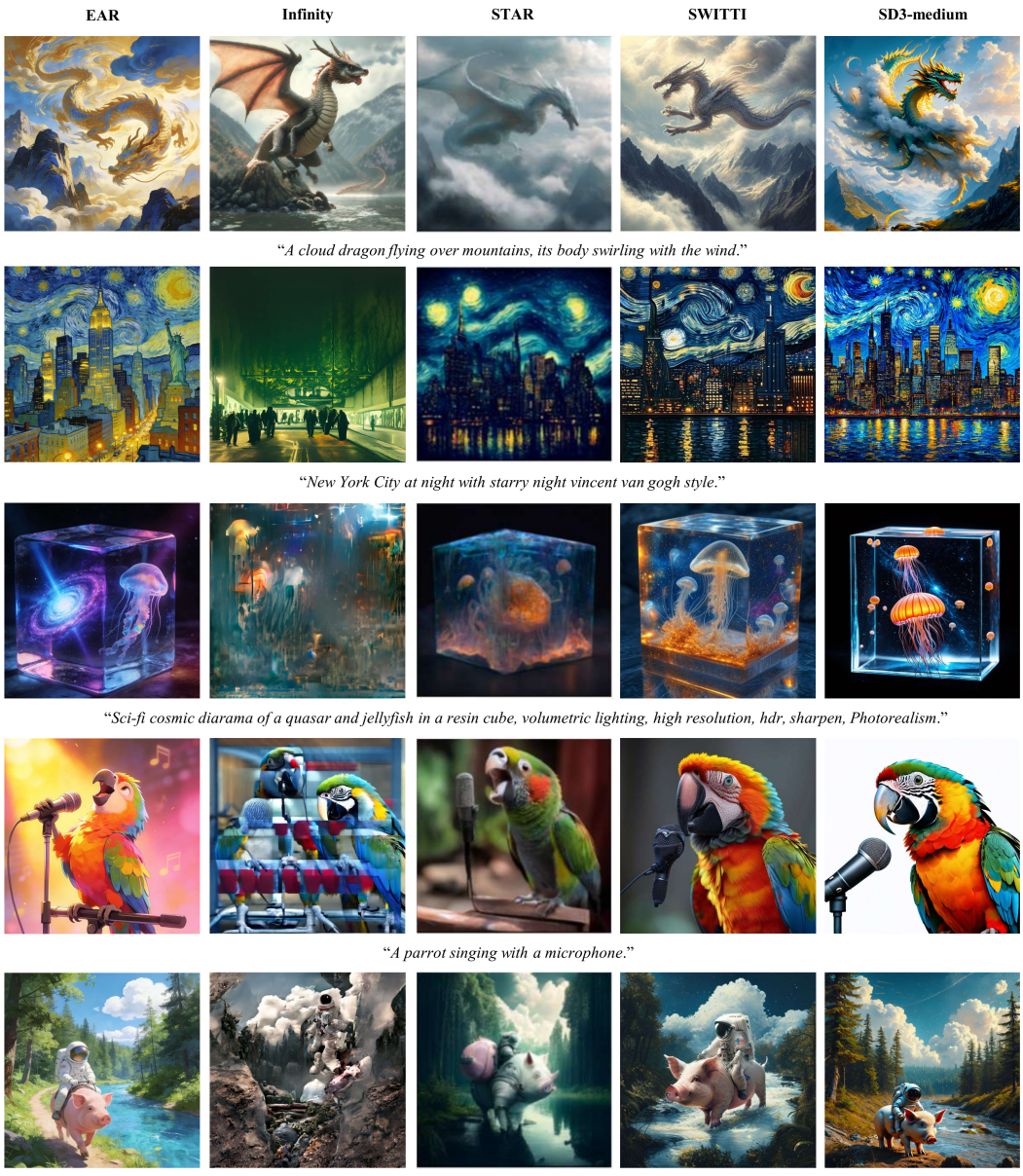

Figure 9: **Qualitative comparison of EAR against the baselines.** All synthesized images are produced with 512 $\times$ 512 resolution.

Table 6: **Subject fidelity and prompt fidelity user preference.** We use majority voting to get the average results.

Model	subject fidelity ↑	prompt fidelity ↑
EAR	38%	47%
Infinity	13%	12%
STAR	11%	10%
SWITTI	15%	13%
SD3-medium	14%	11%
Undecided	9%	7%

A.8 More Qualitative Results

We provide more qualitative comparison of EAR against the baselines involved in this work in Figure 9. From Figure 9, our EAR can generate vivid synthetic images. However, Infinity Han et al. (2025) again produces inferior image generation results, as shown in rows 2 to 5 of Figure 9.

To further evaluate our model's performance in complex multi-object scenarios, we have generated several images using GenEval-style prompts and compared them with state-of-the-art models, including Infinity Han et al. (2025), STAR Ma et al. (2024), SWITTI Voronov et al. (2025), and SD3-medium Esser et al. (2024). The qualitative comparisons are provided in Figure 10. These results demonstrate our model's robust capability in handling complex attribute binding and spatial relationships. Interestingly, Infinity Han et al. (2025) yields poor generation results characterized by distorted shapes and erroneous spatial relationships, which is consistent with the results reported in Table 8.

A.9 More Quantitative Results

Moreover, we provide quantitative ablation studies in Table 7 to correspond to the qualitative examples in Fig. 4. As shown in Table 7, EAR without fine-grained textual guidance results in inferior performance.

Table 7: **Quantitative comparisons of EAR with and without fine-grained textual guidance (denoted as EAR-wot).** All tests are conducted on COCO and MJHQ datasets. Best results are in bold.

Model	COCO 30K				MJHQ 30K				
	PickScore ↑	CLIP ↑	IR ↑	FID ↓	PickScore ↑	CLIP ↑	IR ↑	FID ↓	GenEval ↑
EAR	**0.229**	**0.362**	**1.05**	**11.9**	**0.219**	**0.389**	**0.91**	**5.2**	**0.62**
EAR-wot	0.225	0.357	1.01	12.6	0.215	0.383	0.89	5.6	0.59

Since the subscores of GenEval provides a more granular and comprehensive understanding of our model's performance, we include the GenEval subscores for Infinity Han et al. (2025), STAR Ma et al. (2024), SD3-medium Esser et al. (2024) and our proposed EAR in Table 8.

Table 8: **GenEval scores of different models.** Overall scores have a standard deviation of about 0.01 across random seeds.* result is with prompt rewriting.

Model	Overall	Single object	Two object	Counting	Colors	Position	Attribute binding
EAR	0.62	0.98	0.74	0.67	0.83	0.14	0.34
Infinity	0.73 *	-	0.85 *	-	-	0.49 *	0.57 *
STAR	0.55	0.97	0.70	0.49	0.80	0.13	0.28
SD3-medium	0.65	0.98	0.76	0.69	0.84	0.14	0.35

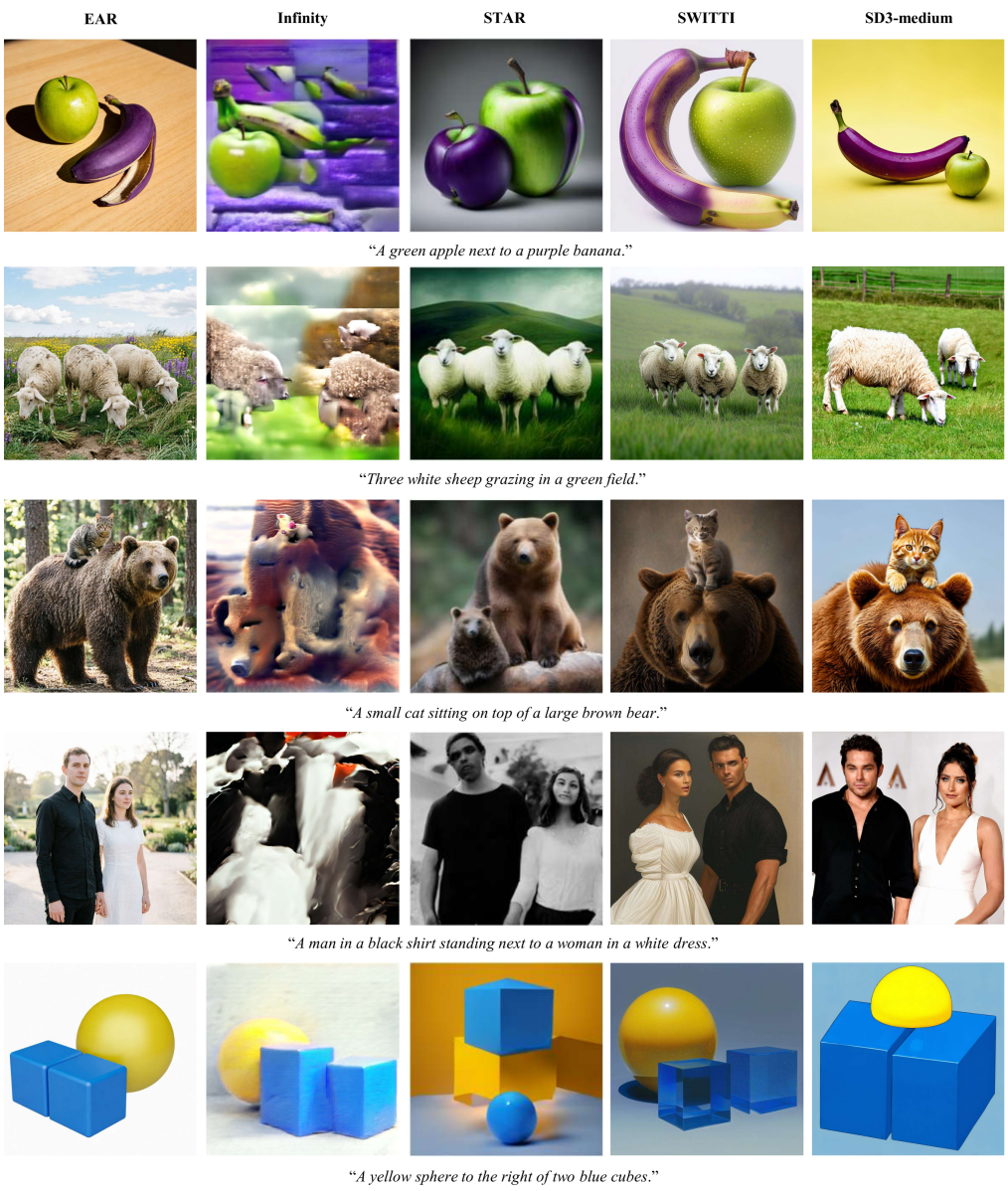

Figure 10: **Qualitative comparison of EAR against the baselines using GenEval-style prompts.** All synthesized images are produced with 512 × 512 resolution.

To comprehensively evaluate the efficiency of our model, we additionally provide a comparison between MaskGIT Chang et al. (2022), MARs Li et al. (2024b), HART Tang et al. (2026), and our proposed EAR in Table 9. It is clear that our proposed EAR achieves the highest throughput.

Table 9: **Comparison of 256 × 256 image generation times.** Best results are in bold.

Model	#Step	Latency(s) ↓	Throughput(img/s) ↑
MaskGIT Chang et al. (2022)	8	0.13	39.49
MARs Li et al. (2024b)	64	18.14	2.93
HART Tang et al. (2026)	10	**0.12**	42.16
EAR	8	0.29	**129.30**

A.10 Practical Insights into Game-Theoretic Modeling

We compare the visual differences of our EAR with variants that employ simpler heuristics (e.g., the locality-aware generation ordering strategy proposed in Zhang et al. (2025), or raster grouping with progressively increasing group sizes), as well as variants based on game-theoretic modeling. The EAR variant using the locality-aware generation ordering strategy is denoted as EAR-loc, while the variant using raster grouping is denoted as EAR-ras.



Figure 11: **Visual comparison between EAR and its variants EAR-loc and EAR-ras.** All images are synthesized at 512 × 512 resolution.

As shown in Figure 11, EAR-loc fails to accurately capture the semantics of the input prompts. For instance, in the first column of Figure 11, EAR-loc cannot correctly place the perfume on a "vanity". In the second column, the attribute "wet" is not correctly synthesized for the leaf by EAR-loc. EAR-ras produces worse results. For example, the perfume in the first column of Figure 11 is

surrounded by a warm golden light, which does not match the input prompt. In the third column, the people exhibit a normal emotion instead of a mysterious look. In contrast, EAR with game-theoretic modeling produces vivid images that are well aligned with the input prompts.

To fully understand the relationship of underlying patterns (spatial, structural, or semantic) to the game-theoretic motivation of our proposed EAR, we aggregated generation orders across 1,000 samples. Statistical analysis reveals that the learned schedules deviate significantly from randomness. Tokens corresponding to the main subject (identified via segmentation masks) are generated, on average, substantially earlier than background tokens. These results, summarized in Table 10, demonstrate that EAR significantly prioritizes semantic subjects over background elements compared to a random grouping schedule. For example, to synthesize foreground regions, EAR uses the first 29.2% of the steps, while random grouping uses 56.2%. Note that experiments are performed on 256×256 images using 20 generation steps. p-values are measured by T-test.

Table 10: **Statistical comparison of generation priority across different image regions (evaluated on 1,000 samples from MS-COCO).** The "Main Subject" regions are identified using pre-trained segmentation masks. A lower average step indicates earlier generation.

Region Type	Avg. Generation Step (1–20) ↓	Avg. Relative Order (%) ↓	p-value (vs. Random)
EAR–Main Subject (Foreground)	5.83	29.2%	<0.001
EAR–Background	14.81	74.0 %	<0.001
Random Grouping–Foreground	11.23	56.2 %	-

A.11 Classifier-Free Guidance

We use classifier-free guidance (CFG) Ho & Salimans (2021) to produce samples of exceptionally high fidelity. Specifically, in our training of EAR, we introduce randomness by occasionally replacing CLIP embeddings with Gaussian noise. During generation, the CFG is represented as follow:

$$e_c = (1 + t) \cdot e_c - t \cdot e_n \tag{15}$$

where e_c is the text conditional embeddings, t is the weight, and e_n is the Gaussian noise embedding. We observe that EAR exhibits a characteristic fidelity-diversity trade-off as the CFG scale varies. Lower scales yield highly diverse images but with weaker adherence to complex text prompts, whereas higher scales lead to over-saturation and diminished diversity. Consequently, we set t=1.5 as the default for our model to balance these factors.

A.12 Training Cost

Both EAR and the standard AR baseline are trained on the same dataset, consisting of approximately 20M LAION image-text pairs and 10M internal images, using the same hardware and optimization settings. Both models follow a two-stage training schedule: 200K iterations at (256×256) resolution with a global batch size of 2,560, followed by 100K iterations at (512×512) resolution with a global batch size of 768. The complete training of EAR requires approximately 2,300 A100 GPU-hours, whereas the standard AR baseline requires approximately 1,900 A100 GPU-hours. Thus, EAR introduces approximately 400 additional A100 GPU-hours, corresponding to a training overhead of 21.1%. Based on our component-wise estimates, the ordering network accounts for approximately 280 A100 GPU-hours, while the MI estimator accounts for approximately 100 A100 GPU-hours. The small difference between their combined cost and the total overhead is due to rounding and shared integration overhead. The MI estimator is used only during training and therefore introduces no inference-time overhead.

A.13 Limitation

A primary limitation of this work is that our theoretical analysis and empirical validation are currently focused on 2D text-to-image synthesis. While the EAR framework is theoretically applicable to any sequence-based generative task, its effectiveness in more complex modalities, such as video generation or 3D synthesis—where temporal or spatial dependencies are more intricate—remains to be explored. Extending the principled generation order to these high-dimensional data types is

a promising direction for future research. Additionally, while EAR achieves a favorable balance between quality and latency, a slight performance gap still exists in certain automated metrics when compared to the largest state-of-the-art models that utilize significantly higher computational budgets.