Discrete Noise Inversion for Next-scale Autoregressive Text-based Image Editing

Quan Dao[1*], Xiaoxiao He[1*], Ligong Han[2†], Ngan Nguyen[3], Amin Heyrani Nobari[4], Han Zhang[5], Faez Ahmed[4], Viet Anh Nguyen[6], and Dimitris Metaxas[1]

[1]Rutgers University [2]Red Hat AI Innovation, MIT-IBM Watson AI Lab
[3]Independent Researcher [4]MIT [5]ReveAI [6]CUHK

Abstract. Visual autoregressive (VAR) models have recently emerged as a promising alternative to diffusion models, achieving competitive performance in text-to-image generation while offering substantially faster inference. Although conditional image generation has been widely studied, training-free prompt-guided image editing remains largely unexplored despite its importance for practical applications. In this paper, we present Visual AutoRegressive Inverse Noise (VARIN), the first noise inversion framework for training-free text-based image editing in visual autoregressive models. VARIN introduces Location-aware Argmax Inversion (LAI), a pseudo-inverse function for the argmax operator that reconstructs the inverse Gumbel noise used during discrete autoregressive sampling. The recovered inverse noise enables accurate reconstruction of the source image while providing controllable guidance for prompt-driven image editing. Extensive experiments demonstrate that VARIN produces edits that closely follow the target prompts while effectively preserving the background and structural details of the source image. These results establish VARIN as a simple, efficient, and practical image editing framework for visual autoregressive models.

Keywords: Pseudo-Inverse Argmax · VAR-based Editing

1 Introduction

In the era of generative AI, diffusion models [9–11, 22, 38, 49] have emerged as dominant methods in image synthesis, outperforming GANs [17] in both unconditional and conditional generation tasks [12], notably in text-to-image generation [44]. This success has enabled diverse practical applications, including personalization [28, 45, 57], 3D generation [39, 61], and prompt-based image editing [3, 20, 21, 25, 34, 37, 65], underscoring their increasing popularity and utility.

In contrast, autoregressive models, traditionally dominant in natural language processing, have only recently gained attraction in visual synthesis. Recent

* Equal contribution.
† Corresponding Author.

Fig. 1: Qualitative performance of VARIN given diverse prompts. VARIN can successfully perform object changing, object adding, background changing and style editing. We can see the the VARIN can perform precisely editing on the editing part of image following target text instruction while preserving the background or unedited part

models such as LlamaGen [50] and MagVit-v2 [66] have optimized image tokenization and transformer architectures, achieving competitive performance with diffusion models. Furthermore, MARS [19] integrates mixtures of experts to train large-scale text-to-image autoregressive models, whereas MAR [30] removes discrete vector quantization by leveraging diffusion-inspired training in continuous space. Despite achieving competitive performance with diffusion, these methods still incur significant inference costs due to dependence on output size, limiting scalability for high-resolution synthesis. To address this issue, [53] proposed a novel next-scale prediction autoregressive (VAR), which predicts the next scale of image instead of next token. Compared to diffusion and traditional autoregressive models, VAR does not only achieve state-of-the-art performance but also possess significant faster inference time than these methods, highlighting the potential future advantages of autoregressive models. Recently, [51] introduced the Hybrid VAR Transformer (HART) diffusion framework, combining visual autoregressive models (VAR) with lightweight diffusion refinement, achieving comparable results to pure diffusion methods but with reduced inference time. This model raises further research opportunities in downstream tasks, including personalization, prompt-based image editing [60], and text-to-3D within VAR-based frameworks.

In this paper, we focus specifically on prompt-guided text-to-image editing within visual autoregressive models (VAR). Given a source image and a text prompt describing desired edits, the task is to edit the source image to align

with the text prompt while preserving background from the original image. We observe that VAR generates the overall structure and layout at initial scales and progressively refines finer details at higher scales. As illustrated in Fig. 2, editing primarily becomes necessary at middle scales, around levels 4 to 6. A straightforward baseline is **Regeneration**, which preserves tokens at initial scales and regenerates remaining scales using the target prompt. However, this approach cannot adequately retains image background.

To address these limitations, we propose **V**isual **A**uto**R**egressive **I**nverse **N**oise **(VARIN)**, an inversion-based training-free editing method designed explicitly for VAR. Inspired conceptually by noise inversion techniques in diffusion models, VARIN identifies a set of editable noises, which is capable of perfectly reconstructing the source image and enabling precise editing. While continuous autoregressive models (e.g., Gaussian-based) allow straightforward inversion through inverse transformation flows [27], discrete VAR models like [51, 53] cannot perform an inversion directly due to the reliance on non-invertible argmax operator during discrete sampling process (Gumbel-max trick). To perform inversion, we require to design a pseudo-inverse function for argmax operator. Recently, DICE [20] proposed an inverse-based editing technique for discrete generative MaskGiT models and introduced a simple pseudo-inverse function for argmax, the one-hot argmax inversion. When applying one-hot argmax inversion for VAR generative models like HART, we find that one-hot argmax inversion produces uncontrollable noise sets which limits precise editing capabilities. Therefore, we propose **Location-aware Argmax Inversion (LAI)**, a pseudo-inverse function, which enhances editing controllability and alignment with target prompts. LAI extracts inverse set of noises to reconstruct source images perfectly and allows adjustable bias towards preserving background details, significantly outperforming simple regeneration baseline. The extracted noises from VARIN allow low source-image bias for editing part of image while keeping high source-image bias for unediting part to preserve background better. Our simple inverse-based VARIN demonstrates good editing quality on par with optimization-based test-time tuning approaches such as Null-Text Inversion [34], without any additional retraining EditAR [35], InstructPix2Pix [4] or complex cross-attention manipulations P2P [21], AREdit [60]. Furthermore, these advanced methods remain complementary and could be combined with our technique in future work for further improvements. Notably, VARIN technique inherits the real-time editing capability from VAR base model, which only cost 0.65s per image editing and significantly outperform diffusion-based editing technique. We summarize our contributions as follows:

1. We introduce **V**isual **A**uto**R**egressive **I**nverse **N**oise (VARIN), a simple yet effective inversion-based editing technique for visual autoregressive (VAR) models. VARIN performs autoregressive noise inversion to extract editable set of noises per scale, enabling controllable image editing without additional training and complex attention manipulation.
2. We propose Location-aware Argmax Inversion (LAI), a pseudo-inverse function of the argmax operator that enables the extraction of inverse noises for

discrete autoregressive sampling. LAI allows perfect reconstruction of the source image while providing controllable source-image bias in the extracted noises, resulting in more precise and stable editing.
3. We empirically validate the effectiveness of VARIN on text-to-image VAR-based generative models, showing that it can generate edits aligned with target prompts while preserving the background and structural details of the source image.

2 Related Work

2.1 Visual autoregressive models.

Autoregressive models have significantly shaped natural language processing (NLP) domain, particularly through their next-token prediction paradigm [1,2,8, 52,54,58,63]. This same paradigm has also proven effective in visual generation: models such as VQVAE [43,56], VQGAN [15,29], DALL-E [41], LlamaGen [50], and MARS [19] tokenize images into sequence of tokens for next-token prediction modelling. While these token-based methods can match diffusion models in image quality, their inference speed often suffers because the number of tokens grows quadratically with image resolution. In visual synthesis domain, recent works rapidly shift toward "next-scale prediction" proposed by VAR [53], where models autoregressively predict multiple scales with different resolution via multiscale residual quantization [29]. VAR [53] and HART [51] exemplify this approach by generating high-quality images in only a few number of function evaluations. Furthermore, VAR is closely related to ImageBART [14], which applies transformers to each denoising step in multinomial diffusion [23]; one can interpret VAR similarly, but using a "blurring-to-deblurring" viewpoint.

2.2 Diffusion editing.

Diffusion-based text-to-image editing has gained popularity for its flexibility and controllability [24, 25, 33, 36]. Many works rely on large-scale pretraining [4, 16, 46, 67, 69] or fine-tuning approaches [13, 18, 47, 70], where either model weights or embeddings are optimized at test time to achieve editing. Null-text Inversion [34] further refines editing control by adjusting "null" embeddings to align the reconstruction path with the source image. Compared to diffusion-based editing technique, VAR-based editing technique like VARIN allows to perform real-time editing within only 0.65s compared to long editing time of diffusion editing technique.

2.3 Diffusion inversion.

A key challenge in diffusion-based editing is extracting an internal representation from the source image so that edits can be made without losing fidelity. Continuous models often invert via deterministic or stochastic processes [7, 25, 31, 32,

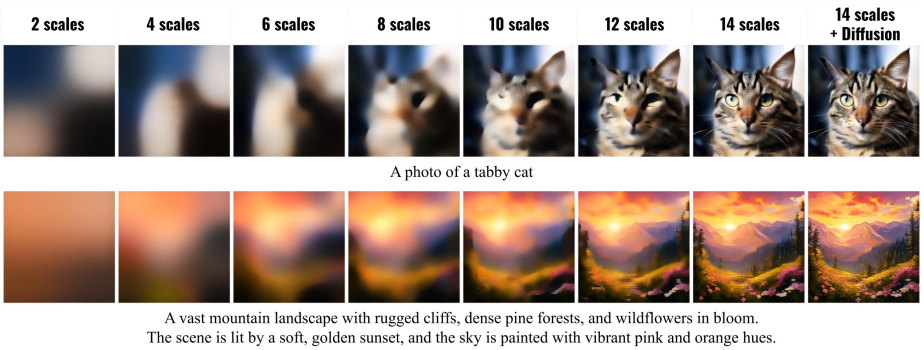

Fig. 2: Visualizations of each scale of the generation process of HART [51]. The features of a cat (top) and the landscape (bottom) are only distinguishable above 6 scales.

48, 64], while recent research addresses discrete diffusion [20] and masked generative models [6]. These methods confirm that "diffusion inversion" can guide image editing by reconstructing the original data from a learned representation. Inspired by these, our work focuses on inverting a visual autoregressive model to enable text-based editing while preserving background.

3 Preliminaries

In this section, we firstly review about the visual autoregressive model in Sec. 3.1. Motivating from SDEdit [33] for diffusion model, Sec. 3.2 introduces simple editing method for VAR called **Regeneration**, where we simply initialize the VAR generative process by a few beginning token scales $r_1, r_2, \ldots, r_t$ extracted from source image using VAR encoder and generate the rest of token scales $r_{t+1}, r_{t+2}, \ldots, r_K$ following target text prompt c_{tgt}. Although this technique produces a fairly good final image that has the same structure as the source image and follows the target text prompt c_{tgt}, they lose background fine details as the final token maps are completely regenerated without any constraint from the source image.

3.1 Visual Autoregressive Model

VAR adopts Variational AutoEncoder (VAE) based on VQ-VAE with $[C] = \{1, \ldots, C\}$ of vocab size C. In training process, given an image $I \in \mathbb{R}^{3\times H\times W}$, the VAR encoder $\mathcal{E}_{\text{VAR}}$ will encode the image into K token scales, $(r_1, r_2, \ldots, r_K) = \mathcal{E}_{\text{VAR}}(I)$, where each token scale r_k has different resolution $h_k \times w_k$ and these resolution increases with the scale k. The VAR-based autoregressive model θ can now be considered as the next scale predictor, which models the following likelihood:

$$p_\theta(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} p_\theta(r_k | r_1, r_2, \ldots, r_{k-1}) \tag{1}$$

where $r_k \in [C]^{h_w \times w_k}$ is token scale, and the sequence $(r_1, r_2, \ldots, r_{k-1})$ is the prefix of r_k. In inference time, we use VAR-based autoregressive to predict the sequence $r_1, r_2, \ldots, r_K$ iteratively. We then pass the generated sequence to the VAR decoder $\mathcal{D}_{\text{VAR}}$ to construct the generative RGB images $I_{\text{gen}} = \mathcal{D}_{\text{VAR}}(r_1, r_2, \ldots, r_K)$.

Unlike traditional autoregressive methods, which flatten the image following a predefined scanning order (i.e. raster scan) then perform the next token prediction, the next scale prediction proposed by VAR [53] strictly follows the autoregressive's behaviour. In VAR framework, each token scale r_k depends only on the previous token scales $r_{<k}$. As a consequence, VAR exhibits faster inference time than diffusion and traditional next-token prediction autoregressive models since they only need to run model for only few scale prediction ($K \approx 14$) instead diffusion and vanilla autoregressive with thousands of model iterations.

3.2 Text-based Image Editing & Regeneration Baseline

Text-based Image Editing: We consider the text-based editing problem for the text-to-image VAR model. Given a source image $I_{\text{src}} \in \mathbb{R}^{3\times H\times W}$ and a target prompt c_{tgt}, we want to edit the source image I_{src} to comply with the prompt c_{tgt}. In our proposed VARIN in Sec. 4, we use inversion algorithm for editing. Therefore, we need to define a corresponding source text c_{src}, which aligns with source image for noise extraction. The source text c_{text} could be written by human or use any image captioning model. The inverse noises are then used for editing process with target prompt c_{tgt} as shown in Sec. 4.3. The image editing is evaluated based on two main criteria, which could be conflicting: the edited image should be **aligned with the target prompt** while still **preserve the unedited part from the source image**.

Baseline Regeneration: Using the VAR encoder, we could obtain the following token scales $(r_1, r_2, \ldots, r_K) = \mathcal{E}_{\text{VAR}}(I_{\text{src}})$. As illustrated in Fig. 2, while the beginning token scale constructs the structure and layout of images, the later token scale adds more finegrain detailed information. For editing with diffusion model, SDEdit [33] starts denoising generative process from middle noise level with guidance condition (i.e. text, canvas) to perform editing, while keeping the overall structure of the source image. Motivating from SDEdit, we could pick a scale index s between 1 and K, then we fix the beginning token scales from $r_1, \ldots, r_s$, and start to regenerate $r_{s+1}, \ldots, r_K$ conditioning on the target prompt c_{tgt}. We name this intuitive editing method **Regeneration**, and outline it in Sec. 4.3.

By keeping the beginning token scales of the source image, the editing output image I_{tgt} could keep the overall structure and layout of the source image. However, it may fail miserably to preserve the finegrained details in background that should not be edited, examples of failures are shown in Fig. 4. To search for a more fine-grained editing control, we propose noise inversion technique and investigate how to control the noise for editing in next section.

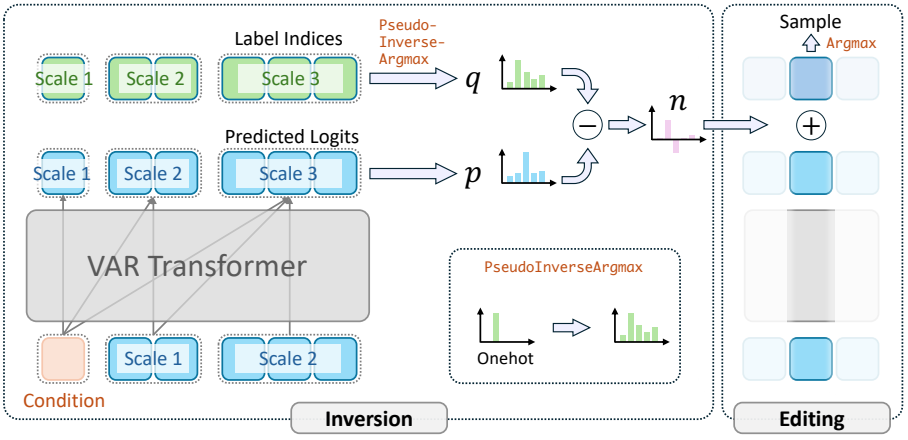

Fig. 3: VARIN pipeline: the prefix token $c_{src} + r_{<k}$ is fed to transformer to get log probability p_k. We then use pseudo inverse-argmax to find the inverse noise n_k from ground truth label r_k and logit p_k. These noise set $n_1, n_2, \ldots, n_K$ is later used for editing control.

4 Inverse Autoregressive Transformation and Editable Inverse Noise

We introduce our technique **V**isual **A**uto**R**egressive **I**nverse **N**oise **(VARIN)** that could preserve the background of source image better. Towards this goal, we discuss the extraction process of noise inversion for an autoregressive model in Sec. 4.1. For discrete space, there is no closed-form solution for inversion due to argmax operator, we therefore propose a pseudo-inverse function for the argmax operator to obtain a set of inverse noises from source image in Sec. 4.2. The overal pipeline is in Fig. 3.

4.1 Inverse Autoregressive Transformations

Given a continuous Gaussian autoregressive model $p_\theta(x_t|x_{<t})$ and a sequence of tokens $\{x_1, x_2, \ldots, x_T\}$, we could apply an inverse transformation in parallel to find the sequence of inverse noises that perfectly reconstruct the sequence [27]. To see this, note that under the Gaussian assumption of $p_\theta(x_t|x_{<t})$, we have $x_t = \mu_t + \sigma_t \odot \epsilon_t$, where μ_t and $\sigma_t > 0$ are the conditional mean and standard deviation of x_t given the history $x_{<t}$. We can obtain ϵ_t by $\epsilon_t = \frac{x_t - \mu_t}{\sigma_t}$. Computing the inverse noise under the Gaussian case could be implemented in parallel due to the independence of ϵ_t and $\epsilon_{t'}$ over different time stamps $t \neq t'$. The above inversion only holds when the noise admits a continuous density. Unfortunately, most of autoregressive generative models for vision including VAR are based on discrete tokens and model the sequence of token with a multinomial distribution. Consequentially, the inversion could no longer be straightforwardly integrated into the VAR pipeline.

We now propose a pseudo-inversion for VAR models that extends the continuous noise inversion to the case of *discrete* tokens. The VAR models use a multinomial distribution, where samples are drawn using the Gumbel-max trick. A Gumbel distribution with mean 0 and scale 1 has a continuous density $\exp\big(-z+\exp(-z)\big)$ for any value z. We consider the discrete inverse autoregressive transformation problem using the Gumbel-max trick as follows:

$$\begin{aligned} &p_t \leftarrow p_\theta(\cdot|r_{<t}) \\ &\text{Find Gumbel noise } n_t \text{ s.t. } \operatorname{argmax}(p_t + n_t) = r_t, \end{aligned} \tag{2}$$

with (unnormalized) log probability $p_t \in \mathbb{R}^{l\times C}$, n_t is the Gumbel noise from Gumbel-max trick, and $r_t \in \mathbb{R}^l$ is the ground truth label with l is number of tokens. As shown in Eq. (2), there is no perfect way to obtain the reverse Gumbel noise from a label and predicted probability, as the Gumbel-max trick involves **an argmax operator**, which is apparently **non-invertible**. Therefore, in the next section, we propose two functions for the argmax operator in Sec. 4.2.

4.2 Pseudo-inverse Argmax

Since $(p_t + n_t)$ represents the Gumbel-perturbed logits, we need to choose the unormalized log probability q_t such that $q_t = p_t + n_t$ and $\arg\max(q_t) = r_t$. Notably, to ensure the editing ability, we better need the n_t to follow the below properties:

- **Prop 1:** n_t needs to be likely sampled from standard Gumbel noise since in line 6 of Algorithm 4, we interpolate n_t with standard Gumbel noise for preserving randomness of generative process.
- **Prop 2:** n_t needs to preserve some bias information from source image. This allows the editing algorithm to preserve the unedited part from source image.

Now, we discuss about how to choose pseudoinverse function for argmax operator. The easiest way is setting $q_t = \text{LogOnehot}(r_t)$, and we call this **Onehot Argmax Inversion (OAI)**. With OAI, we can find the list of inverse noise for perfect reconstruction. However, when using inverse noise OAI for editing, it usually fails to control the editing process. The reason is that the q_t is 0 at the label index and significantly negative at other indices. This results in $n_t = q_t - p_t$ not likely being sampled from a standard Gumbel noise distribution violating **Prop 1** and n_t will be highly biased by source image since q_t is extremely biased by label of token. Therefore, in editing, it is hard to control the editing process with OAI. Furthermore, we notice that the OAI process relies only on ground-truth labels and omits the information of predicted logits p_t from models.

To remedy the above issues, we propose a argmax inversion function called **Location-aware Argmax Inversion (LAI)**. Since $q_t = p_t + n_t$ and n_t is sampled from standard Gumbel noise, q_t should be close to p_t. LAI exploits this property, and it takes p_t as the location of Gumbel-max sampling. This makes q_t closer to p_t, and n_t is more likely sampled from standard Gumbel distribution which is then satisfied **Prop 1**. For the remaining discussion, we

Algorithm 1 Location-aware Argmax Inversion - LAI

Input: tokens $r \in [C]^l$, log prob $p \in \mathbb{R}^{l \times C}$ and preserving term τ

1: $r_{\text{mask}} \leftarrow \text{Onehot}(r, \text{num\_classes} = C)$ ▷ $\in \mathbb{R}^{l \times C}$
2: $l_{\max} \leftarrow \text{Sum}(r_{\text{mask}} \odot p, \text{dim} = -1)$ ▷ $\in \mathbb{R}^{l \times 1}$
3: $q_{\max} \sim \text{Gumbel}(\mu = l_{\max}, \beta = 1)$ ▷ $\in \mathbb{R}^{l \times 1}$
4: $q \sim \text{GumbelTrunc}(\mu = p, \beta = 1, \text{trunc} = q_{\max} - \tau)$ ▷ $\in \mathbb{R}^{l \times C}$, Algorithm 3
5: $q \leftarrow q \odot (1 - r_{\text{mask}}) + q_{\max} \odot r_{\text{mask}}$ ▷ $\in \mathbb{R}^{l \times C}$

Output: q

Algorithm 2 VARIN

Input: Image I_{src}, prompt c_{tgt}
Parameters: τ
Autoregressive Inversion:
1: $(r_1, r_2, \ldots, r_K) \leftarrow \mathcal{E}_{\text{VAR}}(I_{\text{src}})$
2: **for** t from 1 to K **do** ▷ logits
3: $q_t \leftarrow \text{LAI}(r_t, p_t)$ ▷ Algorithm 1
4: $n_t \leftarrow q_t - p_t$
5: **end for**
Return: $(n_1, n_2, \ldots, n_K)$

Algorithm 3 GumbelTrunc

Input: location ϕ, threshold T
1: $u \sim \text{Uniform}(0, I)$
Return: $\phi - \log(\exp(\phi - T) - \log u)$

will omit subscript t to avoid clutter notation. To ensure the argmax condition, for each token i from 1 to l, the ground truth label is $r[i]$, we sample $q[i]$ so that $\text{argmax}(q[i]) = r[i]$. Firstly, we sample the value for $q[i]_{r[i]}$ using the Gumbel-max trick with the predicted location $p[i]_{r[i]}$. Later, we sample other position $q[i]_{\neq r[i]}$ by truncated Gumbel-max trick with corresponding predicted location and truncated value $q[i]_{r[i]} - \tau$ (see Algorithm 3). Please noted that τ is very important parameter to control the unedited part preservation and we shall discuss details of τ below.

Hyperparameter τ in LAI: When $\tau = 0$, the n_t from LAI satisfies **Prop 1** but it still fails to edit while preserving the background. We could explain this effect: For simplicity, we consider $q \in \mathbb{R}^{\mathbb{C}}$ is a vector. When $\lambda = 0$, the q sampled from Gumbel Truncation (Refer to Appendix for further detail) is highly uncertain cause there exist some indices j such that $q_j \approx q_i$, where $i = \text{argmax}_k q_k$. Therefore, during editing, $q^{\text{edit}} = p + (1-\lambda) \cdot g + \lambda \cdot n = p + n + (1-\lambda) \cdot (g - n) = q + (1-\lambda) \cdot (g - n) = q + \epsilon$. Since q is highly sensitive, even a small ϵ can cause a change in the maximum index of q^{edit}, such that $\text{argmax}_k q_k^{\text{edit}} = j$. As a result, preserving the unedited part becomes exceedingly difficult. Therefore, by setting τ, we could let q_t retains bias from source image which provides useful information to n_t and control the sensitivity of Algorithm 4 better. This is also satisfied **Prop 2**.

The efficient implementation is shown in Algorithm 1. LAI always guarantees perfect reconstruction, the same as OAI, but q_t is closer to p_t, and n_t is more like standard Gumbel, which satisfied both above noise properties. We adopt LAI in line 3 in Algorithm 2 to obtain inversed Gumbel noise list to control editing.

Algorithm 4 Editing by VARIN

Input:
$(r_1, r_2, \ldots, r_K) \leftarrow \mathcal{E}_{\text{VAR}}(I_{\text{src}})$
$n_1, n_2, \ldots, n_K \leftarrow \text{VARIN}((r_1, r_2, \ldots, r_K), c_{\text{src}})$
for t from s to K **do** ▷ s is staring scale
  $p_t \leftarrow p_\theta(\cdot|r_{<t}, c_{\text{tgt}})$ ▷ p_t is log probability
  $g \sim \text{Gumbel}(0, I)$
  $q_t = p_t + (1 - \lambda) \cdot g + \lambda n_t$
  $\tilde{r}_t = \text{argmax}(q_t)$
end for
$I_{\text{tgt}} \leftarrow \mathcal{D}_{\text{VAR}}(r_1, \ldots, r_{s-1}, \tilde{r}_s, \ldots, \tilde{r}_K)$
Output: I_{tgt}

4.3 Editing with VARIN

For editing using inversion, similar to Regeneration Editing, we first obtain the token scales using the VAR encoder. We then use Algorithm 2 to collect the list of inverse noise $n_1, n_2, \ldots, n_K$ for each scale. For each t, we get the predicted log probability p_t from autoregressive model p_θ. Instead of sampling using the Gumbel-max trick given p_t as in Sec. 4.3, which uses new Gumbel noise g, we interpolate new noise g with inverse noise n_t by interpolation coefficients λ. The λ is hyperparameter for tuning the editing process.

5 Experiments

In this section, we first describe the evaluation dataset used for the text-based image editing task. In Sec. 5.1, we present the reconstruction performance of our proposed method, VARIN + HART. Then, in Sec. 5.2, we conduct editing experiments to demonstrate that VARIN can effectively perform text-based image editing. Throughout the experiments, we use the HART model [51] as the base model and report quantitative results in Tab. 1 and Tab. 2.

Dataset: To conduct the experiment, we use the Prompt-based Image Editing Benchmark (PIE-Bench) [26] which is a dataset created to evaluate text-to-image (T2I) editing methods. It includes 700 images in 9 different editing scenarios, making it a useful evaluation protocol for testing how well these methods handle text-guided image edits. The dataset contains detailed annotations including source text prompt c_{src} and target prompt c_{tgt}, and variety of tasks for us to thoroughly test and fairly compare our method with other approaches.

5.1 Inversion Reconstruction

In Sec. 5.1, we evaluate the reconstruction performance of VARIN. We firstly adopt Algorithm 2 to generate a set of inverse noises. These inverse noises are then used to reconstruct token scales. Finally, the final token scales are decoded back to images.

Table 1: Inversion reconstruction performance among discrete generative model.

Method		Metric			
Inversion	Model	**PSNR** ↑	**LPIPS**$_{\times 10^3}$ ↓	**MSE**$_{\times 10^4}$ ↓	**SSIM**$_{\times 10^2}$ ↑
DICE	Paella	30.91	39.81	11.07	90.22
VARIN	HART	26.00	70.26	33.20	79.83

Evaluation Metrics. For reconstruction evaluation, we use several image similarity metrics, including Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [68], Mean Squared Error (MSE), and Structural Similarity Index Measure (SSIM) [62], to measure the difference between the source images and the reconstructed images.

Results. In Tab. 1, we compare our inversion method VARIN with DICE [20], since both operate on discrete token distributions. While DICE relies on a discrete diffusion model, our method is built on HART, a discrete visual autoregressive model. In principle, discrete noise inversion in both DICE and VARIN can recover tokens that exactly match the original image tokens, leading to perfect token-level reconstruction. However, the evaluation metrics show a noticeable gap between the reconstructed images and the original source images. This discrepancy arises because both Paella [42] and HART use vector quantization autoencoders, which introduce reconstruction error due to the compression process. As shown in Tab. 1, the VAR-VAE used in HART performs worse than the VQ-VAE employed by Paella in terms of exact pixel-level reconstruction. Nevertheless, although VAR-VAE lags behind VQ-VAE in pixel-to-pixel accuracy, the two models produce visually comparable results. Therefore, this limitation does not hinder their effectiveness in practical applications.

5.2 Text-based Image Editing Performance

This section demonstrates the effectiveness of our editing algorithm VARIN. For both Regeneration and VARIN, we set the starting editing scale to $s = 5$ by default, based on observations from Fig. 2, and set $\tau = 18$ in LAI. Furthermore, we adopt a logarithmic scheduler for λ with respect to scale, where $\lambda = 1$ at scale s and gradually decreases logarithmically to 0 at the final scale 14.

Evaluation Metrics. We evaluate the performance of our proposed editing method across three main aspects: structural similarity, background preservation, and alignment between the edit prompt and the generated image. To assess structural similarity between the original and generated images, we apply the structure distance metric from [55]. For evaluating background preservation outside the editing mask, we use PSNR, LPIPS, MSE, and SSIM. Consistency between the edit prompt and the generated image is measured using the CLIP similarity score [40], computed for the entire image as well as specifically for the regions within the editing mask. Together, these metrics provide a holistic evaluation of our inversion method, assessing its ability to maintain structural

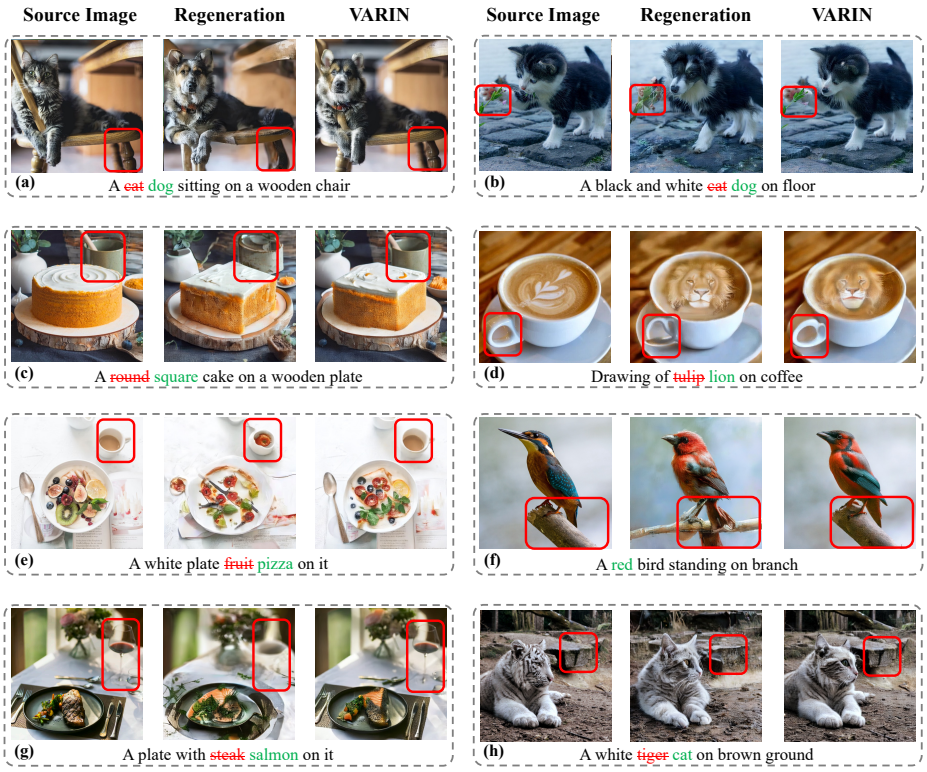

Fig. 4: Comparison between VARIN and the Regeneration baseline. VARIN performs more accurate localized edits while better preserving background details and image structure.

fidelity, preserve background details, and ensure alignment between the editing prompt and the generated image, following the evaluation protocol in [26].

Results. Tab. 2 presents quantitative comparisons between VARIN and representative diffusion-based, autoregressive, and discrete image editing methods. Overall, VARIN achieves a favorable balance between editing accuracy and image preservation. In particular, it consistently ranks among the best methods for structural preservation while maintaining competitive prompt alignment and strong background consistency, demonstrating that autoregressive noise inversion is an effective alternative to diffusion-based editing. We first compare VARIN with recent *training-free* editing approaches including Regeneration, DICE [20], AREdit [60], and DDPM-Inversion [25]. As shown in Tab. 2, VARIN achieves strong performance across multiple metrics while maintaining a favorable balance between editing faithfulness and content preservation.

Among discrete generative editing methods, VARIN demonstrates competitive structural preservation with a Structure Distance of 11.54, comparable to

Fig. 5: Fine-grained editing with VARIN based on Switti [59]. VARIN successfully performs localized edits including facial expression changes, object replacement, and hand gesture modification, while preserving the overall scene structure and background.

DICE (11.34) and substantially better than other autoregressive editing approaches such as EditAR, AREdit. Furthermore, VARIN achieves the highest Whole CLIP within the discrete group (25.59), indicating strong alignment between the generated images and the editing prompts. Compared to DICE [20], VARIN produces stronger semantic alignment in the edited regions (Edited CLIP 22.35 vs. 21.23) while maintaining comparable structural fidelity. DICE exhibits slightly better background preservation, which is expected given the stronger reconstruction capability of Paella tokenizer (see Sec. 5.1). Methodologically, the two approaches differ: DICE relies on discrete diffusion, whereas VARIN bases on next-scale autoregressive framework. Within the category of autoregressive editing methods, VARIN outperforms both methods on most structural and background preservation metrics, while AREdit achieves a slightly higher Edited CLIP (22.77 vs. 22.35).

We next compare VARIN with representative diffusion-based editing methods such as Prompt-to-Prompt [21], Pix2Pix-Zero [41], InstructPix2Pix [4], MasaCtrl [5], and MGIE [16]. VARIN consistently achieves stronger structural preservation than most diffusion-based methods. In particular, it produces substantially lower Structure Distance than Prompt-to-Prompt, Pix2Pix-Zero, and InstructPix2Pix. Compared with stronger diffusion inversion methods such as Null-text Inversion, DDPM-Inversion, and InfEdit, VARIN achieves comparable quantitative performance while avoiding iterative latent optimization during inference. These results demonstrate that autoregressive inversion provides an efficient alternative to optimization-based diffusion editing.

Besides editing quality, VARIN offers significant computational advantages. Most diffusion-based editing methods require iterative denoising and, in many cases, additional optimization during inversion, resulting in relatively high inference costs. In contrast, VARIN performs a single autoregressive inversion followed by next-scale generation, requiring only 0.65 seconds per image on a single NVIDIA A100 GPU. Because VARIN is entirely training-free, it can be

Table 2: Quantitative comparison of text-guided image editing methods on PIE-Bench. We compare VARIN with representative diffusion-based and discrete-based editing methods. VARIN achieves a strong balance between structural preservation, background fidelity, prompt alignment, and inference efficiency.

	Method		Structure	Background Preservation				CLIP Similarity	
	Name	**T2I**	**Distance**$_{\times 10^3}$ ↓	**PSNR** ↑	**LPIPS**$_{\times 10^3}$ ↓	**MSE**$_{\times 10^4}$ ↓	**SSIM**$_{\times 10^2}$ ↑	**Whole** ↑	**Edited** ↑
Continuous	Prompt-to-Prompt	diffusion	69.43	17.87	208.80	219.88	71.14	25.01	22.44
	Negative Prompt	diffusion	16.17	26.21	69.01	39.73	83.40	24.61	21.87
	PnP Inversion	diffusion	**11.65**	27.22	54.55	**32.86**	84.76	25.02	22.10
	Null-text Inversion	diffusion	13.44	27.03	60.67	35.86	84.11	24.75	21.86
	Pix2pix-zero	diffusion	61.68	20.44	172.22	144.12	74.67	22.80	20.54
	MasaCtrl	diffusion	28.38	22.17	106.62	86.97	79.67	23.96	21.16
	InfEdit	diffusion	14.22	**27.52**	**47.98**	34.17	**85.05**	24.89	22.03
	InstructPix2Pix	diffusion	107.43	16.69	271.33	392.22	68.39	23.49	22.20
	MGIE	diffusion	67.41	21.20	142.25	295.11	77.52	24.28	21.79
	DDPM-Inversion	diffusion	22.12	22.66	67.66	53.33	78.95	**26.22**	**23.02**
	RF-Inversion	flow	40.60	20.82	190.00	–	71.92	25.20	22.11
Discrete	DICE	MaskGiT	**11.34**	**27.29**	**52.90**	43.76	**89.79**	23.79	21.23
	EditAR	LlamaGen	39.43	21.32	117.15	130.27	75.13	24.87	21.87
	Regeneration	Hart	25.56	20.45	106.50	95.90	73.31	24.65	21.13
	AREdit	Infinity	30.5	24.19	87.00	–	71.92	25.42	22.77
	VARIN	Hart	11.54	26.47	54.62	38.26	85.27	25.59	22.35

directly applied to pretrained VAR models without additional optimization or fine-tuning.

As demonstrated in Fig. 1, VARIN effectively handles diverse prompts to add objects, change backgrounds, and alter image styles. These edits maintain alignment with the target prompts while preserving essential background details. Additionally, fine-grained editing results using the base model Switti [59] are presented in Fig. 5, illustrating moderate local edits like closing eyes, turning heads, changing objects (e.g., book to iPad), and adjusting hand gestures or leg positions, though substantial pose or structural modifications remain challenging. In the supplementary, we conducted a user study to assess and compare the visualization quality and editing prompt agreement of DDPM-Inv, DICE, and VARIN methods.

6 Conclusion

We propose **VARIN**, a method that enables text-guided image editing for text-to-image visual autoregressive models such as HART. Through extensive experiments, we demonstrate that VARIN can effectively perform text-guided image editing while preserving background content and structural consistency. These capabilities extend HART beyond standard text-to-image generation, making it a more versatile framework for practical image editing applications.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22560–22570 (October 2023)
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
7. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
8. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
9. Dao, Q., Doan, K., Liu, D., Le, T., Metaxas, D.: Improved training technique for latent consistency models. arXiv preprint arXiv:2502.01441 (2025)
10. Dao, Q., Metaxas, D.: Mpdit: Multi-patch global-to-local transformer architecture for efficient flow matching and diffusion model. arXiv preprint arXiv:2603.26357 (2026)
11. Dao, Q., Phung, H., Dao, T.T., Metaxas, D.N., Tran, A.: Self-corrected flow distillation for consistent one-step and few-step image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2654–2662 (2025)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Dong, W., Xue, S., Duan, X., Han, S.: Prompt tuning inversion for text-driven image editing using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7430–7440 (2023)
14. Esser, P., Rombach, R., Blattmann, A., Ommer, B.: Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. Advances in neural information processing systems **34**, 3518–3532 (2021)
15. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
16. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. arXiv preprint arXiv:2309.17102 (2023)

17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
18. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7323–7334 (2023)
19. He, W., Fu, S., Liu, M., Wang, X., Xiao, W., Shu, F., Wang, Y., Zhang, L., Yu, Z., Li, H., et al.: Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. arXiv preprint arXiv:2407.07614 (2024)
20. He, X., Han, L., Dao, Q., Wen, S., Bai, M., Liu, D., Zhang, H., Min, M.R., Juefei-Xu, F., Tan, C., et al.: Dice: Discrete inversion enabling controllable editing for multinomial diffusion and masked generative models. arXiv preprint arXiv:2410.08207 (2024)
21. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems **34**, 12454–12465 (2021)
24. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Chen, S., Cao, L.: Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525 (2024)
25. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12469–12478 (2024)
26. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Direct inversion: Boosting diffusion-based editing with 3 lines of code. arXiv preprint arXiv:2310.01506 (2023)
27. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. Advances in Neural Information Processing Systems **29** (2016)
28. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
29. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11523–11532 (2022)
30. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838 (2024)
31. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
32. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
33. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
34. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)

35. Mu, J., Vasconcelos, N., Wang, X.: Editar: Unified conditional generation with autoregressive models. arXiv preprint arXiv:2501.04699 (2025)
36. Nguyen, T.T., Nguyen, D.A., Tran, A., Pham, C.: Flexedit: Flexible and controllable diffusion-based object-centric image editing. arXiv preprint arXiv:2403.18605 (2024)
37. Pham, C., Dao, Q., Bhosale, M., Tian, Y., Metaxas, D., Doermann, D.: Autoedit: Automatic hyperparameter tuning for image editing. Advances in Neural Information Processing Systems **38**, 6174–6205 (2026)
38. Phung, H., Dao, Q., Dao, T., Phan, H., Metaxas, D., Tran, A.: Dimsum: Diffusion mamba–a scalable and unified spatial-frequency method for image generation. arXiv preprint arXiv:2411.04168 (2024)
39. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
42. Rampas, D., Pernias, P., Aubreville, M.: A novel sampling scheme for text- and image-conditional image synthesis in quantized latent spaces. arXiv preprint arXiv:2211.07292 (2022)
43. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
45. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
46. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024)
47. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8839–8849 (2024)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
49. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
50. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)

51. Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., Han, S.: Hart: Efficient visual generation with hybrid autoregressive transformer. arXiv preprint arXiv:2410.10812 (2024)
52. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
53. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905 (2024)
54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
55. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
56. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
57. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2116–2127 (2023)
58. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
59. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
60. Wang, Y., Guo, L., Li, Z., Huang, J., Wang, P., Wen, B., Wang, J.: Training-free text-guided image editing with visual autoregressive model. arXiv preprint arXiv:2503.23897 (2025), `https://arxiv.org/abs/2503.23897`
61. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
62. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
63. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
64. Wu, C.H., De la Torre, F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 (2022)
65. Wu, C.H., la Torre, F.D.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: ICCV (2023)
66. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)
67. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

69. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9026–9036 (2024)
70. Zhang, Z., Han, L., Ghosh, A., Metaxas, D.N., Ren, J.: Sine: Single image editing with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6027–6037 (2023)