# The Lottery Ticket Hypothesis in Denoising: Towards Semantic-Driven Initialization

Jiafeng Mao[1,2], Xueting Wang[1], and Kiyoharu Aizawa[2]

[1] CyberAgent, Tokyo, Japan
{jiafeng_mao, wang_xueting}@cyberagent.co.jp
[2] The University of Tokyo, Tokyo, Japan
{mao, aizawa}@hal.t.u-tokyo.ac.jp

**Abstract.** Text-to-image diffusion models allow users control over the content of generated images. Still, text-to-image generation occasionally leads to generation failure requiring users to generate dozens of images under the same text prompt before they obtain a satisfying result. We formulate the lottery ticket hypothesis in denoising: randomly initialized Gaussian noise images contain special pixel blocks (winning tickets) that naturally tend to be denoised into specific content independently. The generation failure in standard text-to-image synthesis is caused by the gap between optimal and actual spatial distribution of winning tickets in initial noisy images. To this end, we implement semantic-driven initial image construction creating initial noise from known winning tickets for each concept mentioned in the prompt. We conduct a series of experiments that verify the properties of winning tickets and demonstrate their generalizability across images and prompts. Our results show that aggregating winning tickets into the initial noise image effectively induce the model to generate the specified object at the corresponding location.

**Keywords:** Diffusion Model · Initial Noise

## 1  Introduction

Diffusion models work by progressively denoising initial noise images drawn from a Gaussian distribution. Previous research in this area has primarily focused on the use of text prompts to guide image generation [8,18,27,30–32]. While efforts have been made to enhance the denoising process or explore additional conditions for better control over the generation process [3,9,12,17,22,28], generation failures can still occur. Especially prompts involving multiple objects or precise positional relationships often require users to generate numerous failed images from the same prompt before achieving a satisfactory result.

In this paper, we unveil the existence of migratable pixel blocks in the initial random Gaussian noise that are independently associated with certain concepts. Intriguingly, our discovery mirrors the well-established Lottery Ticket Hypothesis in network pruning [10], suggesting the presence of a valuable subset within randomly initialized entities. Hence, we term our findings in this paper as the lottery ticket in denoising.
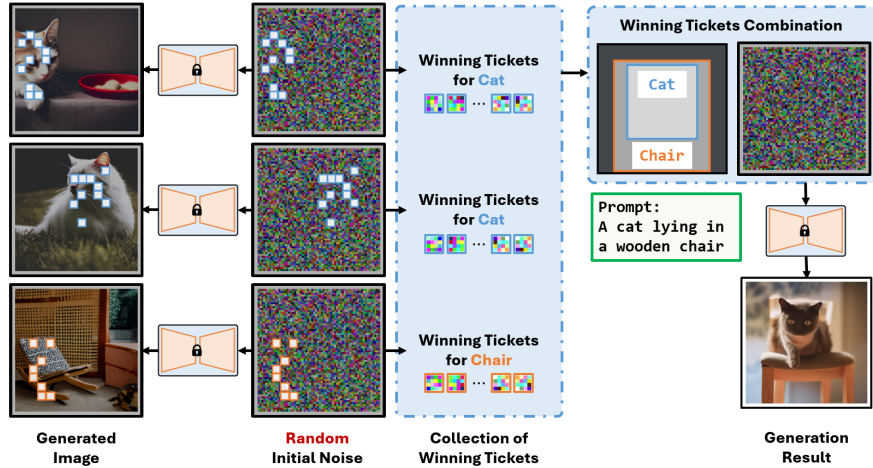
**Fig. 1:** The concept of the winning ticket (special noisy pixel blocks) in denoising. These pixel blocks are naturally susceptible to being denoised to specific concepts. In this paper, we investigate and verify the property of winning tickets. We find that winning tickets collected from different images can be used to create new initial images and show significant advantages under different prompts.

**The Lottery Ticket Hypothesis in Denoising.** *Randomly-initialized noise images contain pixel blocks with initial values that allow them to be used to generate specific concepts.*

Specifically, during a standard generation process, when a user seeks to generate an image under a prompt $\mathcal{P}$ containing an object $c \in \mathcal{P}$, the randomly initialized image will posses special noisy pixel blocks, referred to as the "**winning tickets**", which can readily be denoised into $c$. Typically, these winning tickets are scattered randomly across the initial noise image at various locations. In the final generated image, $c$ will manifest at the location where its winning tickets are most concentrated. This hypothesis explains the reason for the occasional generation failure in standard text-to-image generation synthesis: When there is a significant gap between the spatial distribution of the winning tickets in the random initial image and the spatial distribution required for the content of the prompt description, the generation of the corresponding content fails. Our lottery ticket hypothesis predicts properties of the winning tickets as follows:

– Transplanting winning tickets from their original initial noise to a new noise image does not change the concept $c$ that these winning tickets can be easily denoised to.
– The winning tickets for concept $c$ still win under different prompts containing the same concept $c$.

In this study, we empirically confirm these properties and demonstrate that winning tickets can be extracted from the initial noise and used independently. The establishment of the first property demonstrates that the advantage of these

winning tickets comes from their pixel values rather than from any external condition such as spatial location. The validity of the second property makes it possible to collect winning tickets without knowing the specific generation-time prompt. Combining both properties, we can first collect winning tickets from any prompt including $c$ and afterwards directly transplant them in noise images to generate new prompts containing $c$.

Diffusion models generate realistic images by iteratively refining random Gaussian noise, with current sampling methods typically requiring dozens of denoising steps. In this study, we use a pre-trained diffusion model and identify winning tickets in the random initial noise as early as the first denoising step. Leveraging these identified winning tickets, we pioneer a semantics-driven approach to constructing initial noise, as shown in Fig. 1. Remarkably, **by merely substituting the initial noise image**, without any fine-tuning or disruption to the generation process, we achieve control over the placement of each concept. We also conduct experiments to demonstrate that methods guiding the denoising process yield significantly better results when applied to initial images composed of winning tickets based on our hypothesis.

In our experiments, we create initial noise by combining winning tickets, which does not guarantee that the initial image adheres to the Gaussian distribution. Surprisingly, directly employing pre-trained models on non-Gaussian initial noise still achieves high-quality images and the generated content follows the winning ticket location guidance. This result demonstrates that the diffusion model can tolerate non-Gaussian initial noise, eliminating the requirement for exact adherence to the Gaussian distribution, thereby expanding the operational space for initial image manipulation.

The contributions of this paper can be summarized as follows:

- We introduce the lottery ticket hypothesis in denoising as a novel finding and confirm the inter-image and inter-prompt invariance of winning tickets.
- We verify that the diffusion model can tolerate non-Gaussian initial images and highlight the importance of initialization, providing a new research direction for the interpretation and control of image generation.
- We present a straightforward yet powerful method for identifying and applying winning tickets in denoising, referred to as semantic-driven initial image construction.
- We demonstrate through experiments that initial images constructed by winning tickets consistently exhibit significant advantages, both independently and when combined with methods introducing denoising guidance.

**Implications.** Empirical investigation into the lottery ticket hypothesis in denoising provides valuable insights. By establishing the existence and properties of winning tickets, we hope to leverage this knowledge to:

*Improve fine-grained control.* Given the significant impact of the spatial distribution of winning tickets in the initial image on the generation process, our findings emphasize the importance of considering the initial image when designing methods for achieving fine-grained control.

***For better theoretical understanding of diffusion models.*** The study of winning tickets offers a pathway to deepen our understanding of the denoising process, potentially benefiting research on diffusion model training and sampling.

## 2   Related Work

Diffusion models have emerged as powerful tools for generating realistic images [7, 13, 14, 19, 21, 26, 35]. Originating from pure Gaussian noise, these models iteratively predict the distribution of subsequent images, gradually refining noise samples until high-fidelity images are obtained. Many contemporary methods [8, 11, 18, 31] leverage text prompts projected by CLIP [29] to guide the denoising process. Notably, stable diffusion [32] conducts denoising in a latent space before decoding the latent representation into an image. Moreover, diverse approaches have been explored to introduce conditions enhancing the control over the generation process [2, 4, 15, 17, 22, 25, 28, 34, 37, 39, 43], thereby broadening the utility of diffusion models. These endeavors predominantly focus on augmenting the controllability of diffusion models by refining the generation process.

In the realm of layout-to-image synthesis, generative models are tasked with placing specified objects in user-defined locations within prompts. Certain methodologies require training on annotated images to achieve this objective [1, 5, 6, 16, 41, 42, 44]. Studies have noted that cross-attention maps closely align with the layout and structure of generated images [12], prompting exploration of control strategies based on manipulating these attention maps [3, 9, 23]. BoxDiff [40] propose to update the noise latent during the denoising under the layout guidance. But the initialization of noise is a step that has long been neglected. A recent work [24] explores the generation tendency of initial images, but their discussion is limited to single images and only applicable when the prompt is known.

## 3   Winning Tickets in Denoising

In this section, we delve into the cross-attention mechanism, which facilitates the identification of winning tickets in denoising (Sec 3.1). Building upon our analysis, we present how we discover winning tickets and verify their properties. We construct a collection of lottery tickets, as illustrated in Sec. 3.2, which facilitates winning ticket selection, as detailed in Sec 3.3. Our experiments are conducted using stable diffusion [32].

### 3.1   Cross-Attention Layer and Winning Tickets

**Cross-Attention Layer** The cross-attention maps can serve as a monitor displaying the similarity between each pixel block and each concept mentioned in the text prompt [24]. Stable diffusion augments the U-Net [33] with the cross-attention mechanism [36] to allow for flexible condition. For the text-to-image task, a CLIP is employed to project text $p$ to a representation $\tau_\theta(p)$, which is

subsequently mapped to the intermediate layers of the U-Net through a cross-attention layer as follows,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \tag{1}$$

where $Q = W_Q^{(i)} \cdot z_t$, $K = W_K^{(i)} \cdot \tau_\theta(p)$, $V = W_V^{(i)} \cdot \tau_\theta(p)$. The cross-attention layer computes the similarity between the query $Q$ (i.e., the embedding of the flattened input image $z_t$) and the key $K$ (i.e., the embedding of each word $p^{(i)}$ in the prompt $p$). Hence, the values in the attention map $M^{(i)}$ for a particular word $p^{(i)}$ indicate the corresponding pixel block's similarity with the word $p^{(i)}$. $V$ contains rich semantic features for each word $p^{(i)}$.

**Winning Ticket** Following [24], we partition each random noise image, sized $64 \times 64$, into $4 \times 4$ pixel blocks, resulting in a total of 256 blocks per image. Each attention value on the map $M^{(i)}$ is determined by a $4 \times 4$ pixel block on the noisy image $z_t$ and a word embedding $p^{(i)}$, signifying the resemblance between this block and the concept $p^{(i)}$. Utilizing this attention map, the cross-attention layer reinforces features of $p^{(i)}$ into areas already resembling it, facilitating the denoising of pixel blocks initially exhibiting high cross-attention values for concept $p^{(i)}$. Pixel blocks demonstrating high cross-attention values on the map corresponding to word $p^{(i)}$ are identified as the winning tickets for concept $p^{(i)}$.

Overall, whether a pixel block is considered a winning ticket is determined by its relevance to the concepts. In other words, a pixel block may be considered a winning ticket for concept $c_1$ but not for concept $c_2$. The same pixel block may also become a winning ticket for multiple concepts, as long as it receives high cross-attention values for those concepts. In practical scenarios, when users intend to generate multiple objects, it's imperative to assign distinct winning tickets for each concept mentioned in the text prompt. This means that if a pixel block is identified as a winning ticket for both categories $c_1$ and $c_2$, and the user's desired generation scenario involves both categories, it's crucial to avoid using such ambiguous winning tickets. The reason behind this precaution is to prevent confusion in the denoising process, as utilizing such ambiguous winning tickets may hinder the model's ability to accurately determine whether to denoise the pixel block as category $c_1$ or $c_2$. To find winning tickets with significant tendency, we first collect a large number of noisy pixel blocks and design a selecting method for winning tickets.

### 3.2   Lottery Tickets Collection

In this section, we present a method to collect all pixel blocks (lottery tickets without knowing if they won) along with their category-wise similarity. To construct this database, a list of possible generation categories $\mathcal{L}$ needs to be specified in advance. Specifically, for the given list of possible categories $\mathcal{L}$, we enumerate all possible combinations of two different objects, $\{c_1, c_2\}$, to construct a list of prompts $\mathcal{L}_p$. The prompts take the form of [a $c_1$ and a $c_2$].
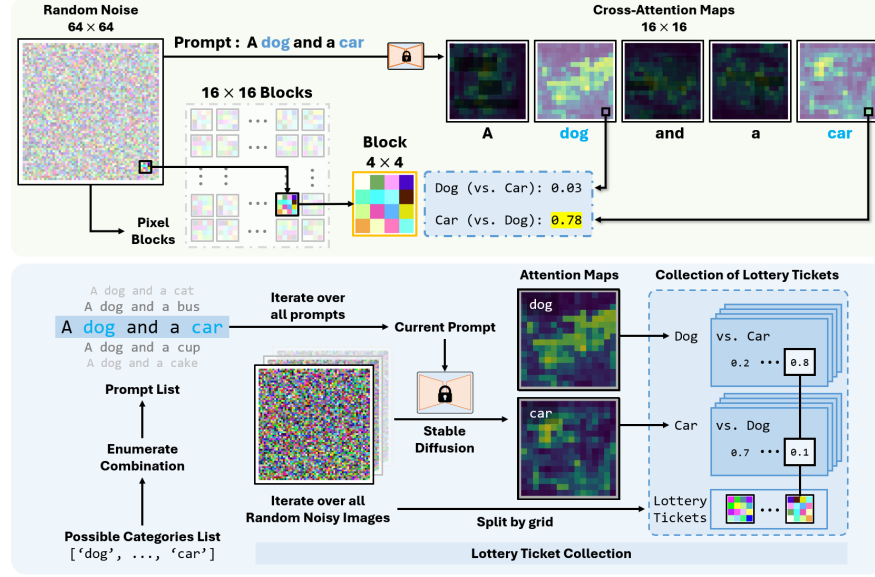
**Fig. 2:** We create a collection containing a large number of pixel blocks and their scores on each possible category. The scores are extracted from the cross-attention maps calculated by the pre-trained stable diffusion. We will select winning tickets from this collection in the next phase.

In practical, we find that when the prompt mentions only one object (e.g., a photograph of a $c$), the pixel blocks with high attention values for the word $c$ are winning tickets for "foreground" rather than winning tickets for the object $c$, due to the absence of competition between different concepts. Therefore, in our experiments, we utilize contrastive prompts, i.e., prompts composed of a pair of objects (e.g., [a $c_1$ and a $c_2$]), and perform a single step of denoising through the U-Net in the model. We collect the similarity information between tickets and these two concepts according to the cross-attention maps, as shown in the upper part of Fig. 2. Since the attention values of blocks are obtained through soft-max over words, obtaining a high score for word $c_1$ implies that this pixel block is a winning ticket for word $c_1$ over word $c_2$. We believe that such a design enables the screening of winning tickets with significant semantic information.

As we illustrated in Sec. 3.1, whether a pixel block is considered a winning ticket is determined by its relevance to the corresponding concept. In each generation under the prompt containing different categories, different pixel blocks become winning tickets. For simplicity, we collect all lottery tickets from $\mathcal{N}$ noisy images $\mathcal{Z} = \{z_1, z_2, ..., z_\mathcal{N}\}$ sampled from the Gaussian distribution into a collection of lottery tickets. As shown in Fig. 2, for each noisy image $z_i \in \mathcal{Z}$, we iterate through all prompts $p_j \in \mathcal{L}_p$. Utilizing the combination $\{z_i, p_j\}$, we only compute the cross-attention maps $\{\mathcal{M}_{c_1}, \mathcal{M}_{c_2}\}$ for the first step of denoising, where $\{c_1, c_1\} \subset p_j$. The final cross-attention values are calculated via the soft-

max function across all words within the prompt [36]. Each pixel block $b_i^k \in z_i$ is logged into the collection, along with its score on $\{\mathcal{M}_{c_1}, \mathcal{M}_{c_2}\}$ in the entry of $c_1$ (vs. $c_2$) and $c_2$ (vs. $c_1$), respectively. Such an operation is performed on all combinations $\{z, p \mid z \in \mathcal{Z}, p \in \mathcal{L}_p\}$. For simplicity, we use $\mathcal{S}_{1,(2)}$ to denote the $\mathcal{S}$core entry of $c_1$ (vs. $c_2$) in the collection. A high score in $\mathcal{S}_{1,(2)}$ implies that the corresponding lottery ticket can be a winning ticket for $c_1$ when $c_1$ and $c_2$ both appear in the prompt. For each lottery ticket $b$ in the collection, we calculate an average score for each category $c_i$ by averaging across all combinations of $c_i$ (vs. $c_j$), where $j \neq i$, as follows,

$$\text{Average Score}_i = \frac{1}{\mathcal{N}_c - 1} \sum_{j \neq i} \mathcal{S}_{i,(j)} \tag{2}$$

where $\mathcal{N}_c$ indicates the number of possible categories.

### 3.3 Winning Tickets Selection

We select winning tickets based on our constructed collection of lottery tickets, which contains a diverse set of pixel blocks, each associated with scores for different category pairs. A high score in $c_1$ (vs. $c_2$) indicates that the corresponding pixel block is well-suited for generating $c_1$ when $c_2$ is also present in the prompt. we can select winning tickets for any list of categories $\mathcal{C} = \{c_i, c_2, ..., c_n\} \subseteq \mathcal{L}$. Specifically, we select winning tickets $\mathcal{B}_i$ for $c_i$ using the following criterion:

$$\mathcal{B}_i = \{b \mid \forall c_j \in \mathcal{C}, j \neq i, \mathcal{S}_{i,(j)}(b) > \mathcal{T}_{obj}\}, \tag{3}$$

where $\mathcal{S}_{i,(j)}(b)$ denotes the score in the entry of $c_i$ (vs.$c_j$) corresponding to the lottery ticket $b$, and $\mathcal{T}_{obj}$ denotes a predefined threshold.

Although we collect scores by category pairs, we experimentally demonstrate that using **prompts containing more than two objects** during the generation does not affect the property of the winning tickets we collect, as shown in Sec. 4.2. In the case where the given list of categories $\mathcal{C}$ only contains a single category $c_i$, we extend the utilization of the average score to select winning tickets specifically tailored for that singular object $c_i$, as follows,

$$\mathcal{B}_i = \{t \mid \text{Average Score}_i(b) > \mathcal{T}_{obj}\}, \tag{4}$$

where Average Score$_i(b)$ denotes the average score of the pixel block $b$ on the category $c_i$. We select pixel blocks with low average scores for all categories $c_i \in \mathcal{C}$ to obtain winning tickets for the background.

$$\mathcal{B}_{bg} = \{b \mid \forall c_i \in \mathcal{C}, \text{Average Score}_i(b) < \mathcal{T}_{bg}\}, \tag{5}$$

where $\mathcal{T}_{bg}$ is a predefined threshold.
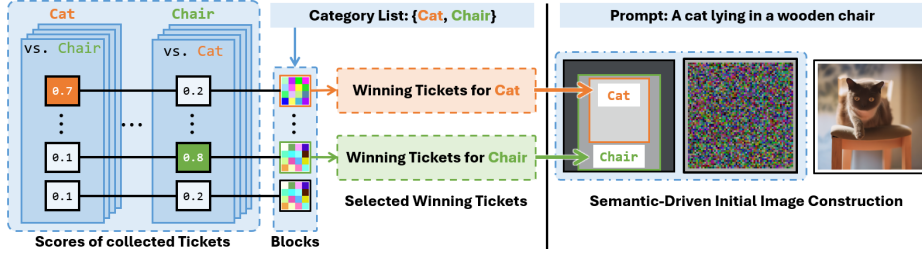
**Fig. 3:** The collection contains scores for all lottery tickets, contrasting each category against others ($c_i$ (vs. $c_j$)). These scores serve as critical metrics used in the selection.

## 4   Property Verification of Winning Tickets

We present the most straightforward application of the collected winning tickets in Sec. 4.1, i.e., constructing regions in the initial noise by the winning tickets to induce the model to generate the corresponding concepts in these regions. We present two distinct usages and their results in Sec. 4.2 and Sec. 4.3, which lead to interesting insights and analysis, as detailed in Sec. 4.4.

### 4.1   Semantic-Driven Initial Image Construction

Following the process we introduced in Sec. 3, given a category list $\mathcal{L}$, we can obtain winning tickets for each concept $c \in \mathcal{L}$. This process has two distinct usage scenarios as follows:

- On the fly: After the user has given the target generation prompt, our method can immediately create a collection based on the list $\mathcal{L}$ of categories extracted from the prompt and select the winning tickets from the collection before generating the image (Sec. 4.2).
- In advance: Our method can create a collection based on a list $\mathcal{L}$ containing a large number of possible categories in advance, so that winning tickets can be selected directly from this collection when generating images.(Sec. 4.3).

We present a straightforward strategy of ultilizing these selected winning tickets. Specifically, for a prompt and specified locations of each object given by the user, we construct the corresponding area in the initial image by the winning tickets for the corresponding concept, as illustrated in Fig. 3. We randomly select the number of pixel blocks needed to construct the specified region from the filtered winning tickets and randomly fill the corresponding region with these selected winning tickets. We refer to this process as **Sem**antic-driven **I**nitialization (SemI) as we construct the initial noise using winning tickets with recognized semantic information. After that, we use the pre-trained diffusion model to denoise our constructed initial noise normally. When the number of winning tickets available is fewer than the number needed, we randomly reuse the winning tickets. We utilize the default configuration of SD [32] v1.4 for the generation. In this setting, the model undergoes 50 denoising sampling steps using DDIM [35].
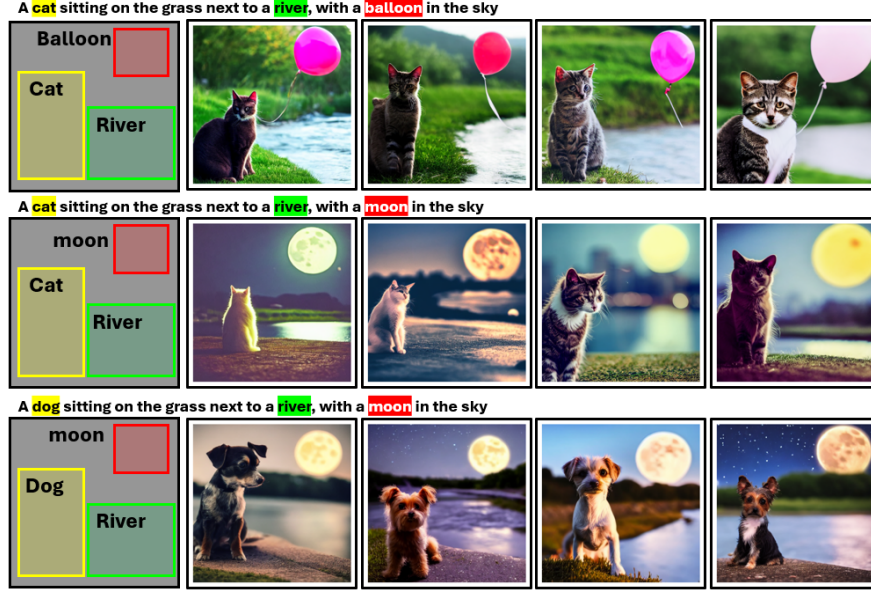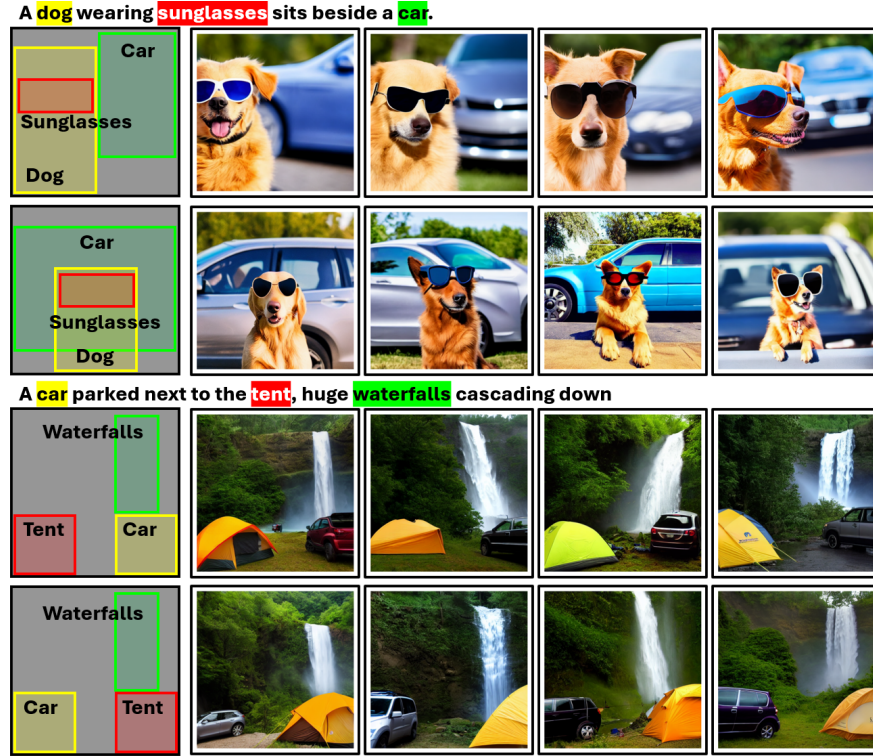
**Fig. 4:** Quality samples in our on-the-fly experiments, under the same specified regions and different prompts. By merely replacing the constructed initial noise image, without interfering with the generation process, the model spontaneously generates the specified concepts in regions consisting of the winning tickets corresponding to these concepts after dozens of denoising steps.

## 4.2   Collection On the Fly

In this section, we manually create prompts containing multiple objects and manually specifying the locations of the individual objects to confirm the effectiveness of winning tickets in application. After we are given the prompt $p$, we use the set of objects contained in the prompt as the possible category list $\mathcal{L}$ and select the winning tickets from multiple random noise images. As shown in Fig. 4, and 5, selected winning tickets proficiently direct the generation of the three specified objects. Moreover, if descriptive words related to the background, such as sky, lake, etc., are included in the categories list during the collection creation, we can also find winning tickets for these background elements. In Fig. 6, the top and bottom halves utilize identical prompts, differing only in the designated background regions. This difference is prominently reflected in the denoising results, demonstrating the power of winning tickets.

The collection creation incurs a time cost. Let $\mathcal{N}_p$ denote the number of possible category pairs and the number of initial images is $\mathcal{N}$, the time cost of creating the collection is $\mathcal{N}_p \cdot \mathcal{N} \cdot t$, where $t$ represents the time required for a single denoising step. For stable diffusion v1.4 on an NVIDIA V100, $t$ is approximately 0.15 seconds. For instance, when the given prompt includes 3 objects, creating a collection with $\mathcal{N} = 10$ takes roughly 5 seconds, and $\mathcal{N} = 100$

A **dog** wearing **sunglasses** sits beside a **car**.



A **car** parked next to the **tent**, huge **waterfalls** cascading down



**Fig. 5:** Results using the same prompt and different specified regions. The region composed of the winning tickets effectively induces the model to generate objects in it.

requires approximately 1 minute. The time cost of the winning ticket selection and the construction of the initial image are negligible.

### 4.3 Collection in Advance

To verify the behavior of winning tickets in the presence of a large number of object combinations, we conducted an experiment where winning tickets were collected in advance against a large number of object pairs. We conduct this experiment on the MS COCO 2017 dataset [20], which contains both image captions and object annotations. The captions of images can be used as prompts, and the object annotations can be used as specified regions of objects. We select 6000 samples whose bounding box categories are all mentioned in their image captions, where 3000 contain one object, and the other 3000 contain two. There are a total of about 700 possible category combinations in the samples.

**Control Strength Evaluation**  To quantitatively evaluate the effectiveness of winning tickets, we use an object detector [38] to perform object detection

**Fig. 6:** Results using the same prompt and different specified background regions.

on generated images and obtain the categories and detected bounding boxes of generated objects [23]. Then the consistency of the detection results with the layout guidance can represent the effectiveness of our controlling. For each object, the Intersection over Union (IoU) between its layout guidance bounding box (i.e., object annotation in the dataset) and detected bounding box is calculated. If the IoU is larger than 0.5, it is regarded as a successful control, and the success rate is represented by $R_{\mathrm{suc}}$. To evaluate the effectiveness of objects of different sizes, all objects are categorized into several subsets according to their size. The subset $s$ contains objects with an area less than $150^2$ pixels, the subset $l$ contains all objects with an area greater than $300^2$ pixels, while the other objects are channeled into subset $m$.

**Implementation** Stable diffusion pre-trained for the text-to-image task does not receive any layout guidance, but may by chance generate the specified objects in the specified regions. For experiments on other methods, the model receives both prompts and layout guidance and produces corresponding images. Our method uses the same pre-trained model to create our collection. The results of our method reported in Tab. 1 are obtained under the default hyper-parameters of $\{\mathcal{N}, \mathcal{T}_{obj}, \mathcal{T}_{bg}\} = \{100, 0.5, 0.1\}$. An analysis of hyper-parameters is provided in Sec. 4.3. In this experiment, we create a collection based on all possible pairs of categories ($\sim 700$) in the dataset. This results in approximately 3 hours to create the collection of lottery tickets.

**Combination with related methods** Some training-free mask-based methods [3, 23] are proposed for the layout-to-image synthesis task. We compare our method to them, and also since these methods focus on providing additional guidance merely through adding masks on cross-attention maps during the denoising process, they are fully orthogonal to what is discussed in this paper. We

| Methods | IoU ↑ | IoU$_s$ ↑ | IoU$_m$ ↑ | IoU$_l$ ↑ | $R_{\mathrm{suc}}\%$ ↑ | $R_{\mathrm{suc}}^{s}\%$ ↑ | $R_{\mathrm{suc}}^{m}\%$ ↑ | $R_{\mathrm{suc}}^{l}\%$ ↑ |
|---|---|---|---|---|---|---|---|---|
| SD [32] | 0.19 | 0.05 | 0.21 | 0.38 | 12.14 | 0.50 | 8.11 | 39.41 |
| Paint [3] | 0.24 | 0.08 | 0.27 | 0.46 | 17.56 | 1.16 | 14.05 | 52.34 |
| Soft [23] | 0.27 | 0.10 | 0.31 | 0.48 | 22.31 | 2.95 | 22.38 | 56.24 |
| Swap [24] | 0.26 | 0.09 | 0.32 | 0.46 | 21.58 | 2.26 | 22.75 | 54.21 |
| SemI$_{\mathrm{(ours)}}$ | 0.26 | 0.10 | 0.32 | 0.45 | 21.94 | 2.40 | 24.68 | 52.31 |
| SemI + Soft [23] | 0.31 | 0.14 | 0.38 | 0.50 | 28.73 | 5.45 | 33.49 | 62.34 |
| SemI + Swap [24] | 0.28 | 0.11 | 0.34 | 0.46 | 23.73 | 3.38 | 27.25 | 54.21 |
| SemI + [23] & [24] | **0.32** | **0.15** | **0.39** | **0.52** | **29.29** | **5.73** | **34.42** | **62.74** |

**Table 1:** Constructing initial noise solely through the utilization of winning tickets achieves a control effectiveness comparable to existing layout-to-image methods, and incorporating winning tickets alongside them significantly enhances their performance.

try to use our method in conjunction with these methods to confirm that initialization using winning tickets improves the performance of these related methods. A recent work [24] proposes to manipulate initial noise inside only a single noise image but suffers from the limited number of winning tickets. Therefore, we also try to improve its performance by combining them with ours.

**Results** As shown in Tab. 1, our method achieves better performance than the Paint [3] and Swap [24], and comparable performance with Soft [23], in terms of control effectiveness, and combining with ours significantly boosts the performance of all other methods. This experimental result exemplifies the superiority of using winning tickets to construct the initial noise.

**Hyper Parameters** We conducted a quantitative sensitivity analysis by varying parameters within the set $\{\mathcal{N}, \mathcal{T}_{obj}, \mathcal{T}_{bg}\}$ and observed the resulting changes, as depicted in Tab. 2. Larger values of $\mathcal{T}_{obj}$ and smaller values of $\mathcal{T}bg$ improve the control over small and medium-sized objects. However, this heightened control leads the value distribution of winning tickets to severely deviate from Gaussian distribution, thereby leading to worse performance for control of large-size objects, where more winning tickets are utilized to construct the initial noise.

Our strategy of reusing winning tickets, as illustrated in Sec. 4.1, is influenced by $\mathcal{N}$. A smaller $\mathcal{N}$ implies fewer pixel blocks in the collection, causing the constructed image to be more likely to contain duplicate pixel blocks, especially for control of large-size objects. Thus, a larger $\mathcal{N}$ yields better image quality.

## 4.4   Discussion

The successful generation and the ability to exert control through utilizing winning tickets are enlightening for **three significant reasons** as follows:
**Tolerance for non-Gaussian Initial Images** All of the initial images used for denoising in our methods are constructed and do not adhere to the Gaussian

| Parameters | IoU $\uparrow$ | IoU$_s$ $\uparrow$ | IoU$_m$ $\uparrow$ | IoU$_l$ $\uparrow$ | $R_{\text{suc}}\%$ $\uparrow$ | $R_{\text{suc}}^s\%$ $\uparrow$ | $R_{\text{suc}}^m\%$ $\uparrow$ | $R_{\text{suc}}^l\%$ $\uparrow$ |
|---|---|---|---|---|---|---|---|---|
| Default | 0.26 | 0.10 | 0.32 | 0.45 | 21.94 | 2.40 | 24.68 | 52.31 |
| $\mathcal{N} = 10$ | 0.25 | 0.08 | 0.31 | 0.44 | 21.76 | 2.10 | 25.00 | 51.45 |
| $\mathcal{N} = 50$ | 0.26 | 0.09 | 0.32 | 0.45 | 22.98 | 3.05 | 26.31 | 53.01 |
| $\mathcal{T}_{obj} = 0.3$ | 0.25 | 0.08 | 0.31 | 0.45 | 20.32 | 1.76 | 21.91 | 50.90 |
| $\mathcal{T}_{obj} = 0.7$ | 0.25 | 0.10 | 0.32 | 0.41 | 21.32 | 3.05 | 25.12 | 47.59 |
| $\mathcal{T}_{bg} = 0.05$ | 0.26 | 0.10 | 0.32 | 0.45 | 22.36 | 3.02 | 24.97 | 52.56 |
| $\mathcal{T}_{bg} = 0.2$ | 0.25 | 0.08 | 0.31 | 0.45 | 21.19 | 2.10 | 22.55 | 53.11 |

**Table 2:** Quantitative sensitive analysis over all hyper-parameters. The default setting is $\{\mathcal{N}, \mathcal{T}_{obj}, \mathcal{T}_{bg}\} = \{100, 0.5, 0.1\}$.
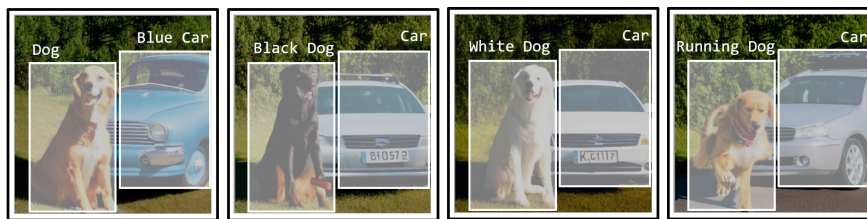


**Fig. 7:** Versatility of winning tickets over different prompts containing the same objects. These images are generated from the same constructed noise and **different prompts**. The tickets constructing each area are selected according to their scores on the simple prompt **a dog and a car**. The generation tendency of those tickets remains unchanged even with different descriptions, such as **black** or **running** dog.

distribution. According to our results (Fig. 4, 5 and Tab. 1), most of the images generated from non-normal initial images still exhibit high quality while being effectively guided. However, the diffusion model can tolerate non-Gaussian initial images to a certain extent, but the generation still fails when the initial image deviates extremely from the normal distribution.

**Versatility of Winning Tickets over Prompt** In our experiment, we create a collection in advance by simply accepting a list of possible categories, without knowing the specific prompts that the generation uses. We adopt a simplistic prompt format (e.g. `[a dog and a cat]`) to select the winning tickets. According to our evaluation, winning tickets evaluated under such simple prompts can generate objects in various prompts efficiently. Our results reveal that a winning ticket has considerable versatility between different prompts containing the same objects. Winning tickets of `dog` under our simple prompt `[a dog and a car]` demonstrate applicability across various prompt descriptions, such as `running dog` or `black dog`, as shown in Fig. 7.

**Versatility of Winning Tickets over Image** Our results confirm the versatility of winning tickets across different initial images. Whether a pixel block can be a winning ticket is determined by its initial pixel values. winning tickets from one image will win again when transferred to another image.
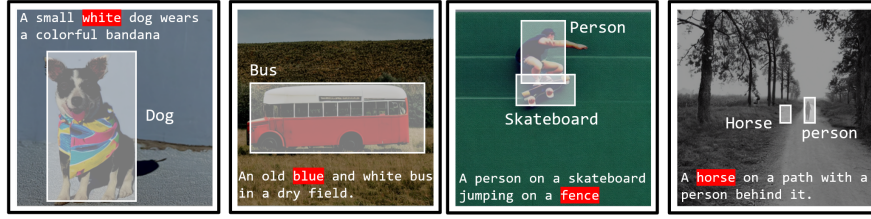
**Fig. 8:** Typical failure cases in our experiment. Each image is generated from a constructed noise according to the layout guidance denoted by the white layout guidance.

### 4.5   Limitations and Future Works

For clarity of explanation, we use typical failure cases to illustrate the limitations of our approach of utilizing winning tickets, as shown in Fig. 8. In the first two images, both the dog and the bus have the wrong colors. The winning ticket selection in our method is based solely on the names of the categories. Occasionally, these selected winning tickets may exhibit a tendency to generate certain random properties (e.g. `red` bus), and such tendencies overcome the prompt during the generation. This can probably be addressed by a finely designed method of selecting winning tickets considering other words in the prompt.

The position control of the third image is successful, but the quality of this image is compromised and the background generation fails. As a small number of the initial noise constructed by winning tickets deviates extremely from the normal distribution, images generated from them carry certain unnatural features. Recalibrating the constructed initial image to adhere to a normal distribution may alleviate this issue. The last image contains two very small objects, and our method only succeeds in generating the person. Although our method significantly outperforms other methods on control over small objects, when the guidance box is extremely small, the generation is still challenging.

## 5   Conclusion

This work introduces and validates the Lottery Ticket Hypothesis in Denoising for text-to-image generative diffusion models, revealing that certain pixel blocks within initial noise images—termed *winning tickets*—possess inherent predispositions towards generating specific content. Our findings illuminate a previously underexplored aspect of the denoising process and propose a novel, semantic-driven approach for initial image construction that significantly enhances control over the generated imagery. By empirically demonstrating the existence, properties, and applicability of winning tickets across various images and prompts, we have opened new pathways for improving the precision and flexibility of image generation. This research not only advances our understanding of diffusion models but also sets the stage for further explorations into optimizing generative processes for more deterministic and customizable outputs.

# References

1. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: CVPR. pp. 18370–18380 (2023)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
5. Chen, K., Xie, E., Chen, Z., Hong, L., Li, Z., Yeung, D.Y.: Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. arXiv (2023)
6. Cheng, J., Liang, X., Shi, X., He, T., Xiao, T., Li, M.: Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. arXiv (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021)
8. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. NeurIPS **34**, 19822–19835 (2021)
9. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. ICLR (2023)
10. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: International Conference on Learning Representations (2018)
11. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: ECCV. pp. 89–106. Springer (2022)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. ICLR (2023)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop (2021)
15. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. ICML (2023)
16. Jia, C., Luo, M., Dang, Z., Dai, G., Chang, X., Wang, M., Wang, J.: Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. arXiv (2023)
17. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. CVPR (2023)
18. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR. pp. 2426–2435 (2022)
19. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR. pp. 22511–22521 (2023)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)

21. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: ICLR (2022)
22. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: ECCV. pp. 423–439. Springer (2022)
23. Mao, J., Wang, X.: Training-free location-aware text-to-image synthesis. ICIP (2023)
24. Mao, J., Wang, X., Aizawa, K.: Guided image synthesis via initial image editing in diffusion model. ACM MM (2023)
25. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv e-prints pp. arXiv–2302 (2023)
26. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171. PMLR (2021)
27. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML. pp. 16784–16804. PMLR (2022)
28. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: NeurIPS Datasets and Benchmarks Track (Round 1) (2021)
29. 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: ICML. pp. 8748–8763. PMLR (2021)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
31. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831. PMLR (2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
34. 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: CVPR. pp. 22500–22510 (2023)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
37. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. SIGGRAPH (2023)
38. WANG, C.Y., YEH, I.H., LIAO, H.Y.M.: You only learn one representation: Unified network for multiple tasks. Journal of Information Science and Engineering **39**(3), 691–709 (2023)
39. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. In: arXiv (2022)
40. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
41. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: CVPR. pp. 14256–14266 (2023)

42. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: CVPR. pp. 14246–14255 (2023)
43. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. ICCV (2023)
44. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: CVPR. pp. 22490–22499 (2023)