

RankT2I: A Submodular Framework for Discovering Interpretable and Diverse Semantics in Text-to-Image Models

Ritika Allada¹ and Pinar Yanardag¹

Virginia Tech, Blacksburg, VA, USA
{ritika88, pinary}@vt.edu

Abstract. Recent advances in text-to-image (T2I) models have revolutionized the field of image generation and editing. However, identifying semantics that a T2I model can successfully edit in an image continues to be a challenging task. Most existing approaches require users to manually specify semantics to modify a particular image, a time-consuming process that often involves extensive trial and error. In this paper, we present RankT2I, a novel, training-free, and model-agnostic framework that automates the discovery of editable semantics in diffusion and FLUX-based models. Given a visual domain, we first utilize a multimodal vision-language model to gather a broad set of candidate semantics. We then frame semantic discovery as a set selection problem and use a submodular objective to identify semantics that are relevant, editable, and diverse. Our method helps users efficiently identify a wide range of semantics for text-to-image editing models across several domains while outperforming existing methods.

Keywords: Semantic Discovery · Text-to-Image Models · Image Editing

1 Introduction

Text-to-image (T2I) models [51; 54; 57; 58; 60; 75; 80] have rapidly advanced the field of image generation and editing. In practice, these models are used to create stylistic changes [9; 23; 46] or add or remove fine-grained features in an image [4; 53; 86]. Some applications of generative image editing include improving the lighting of photos [7], personalizing digital images [17], assisting users with virtual try-ons in the fashion industry [11; 36], and more.

Recent works have focused on developing methods to improve the controllability of fine-grained image editing, such as manipulating specific latent features or using text prompts to guide the editing process [16; 31; 32; 52; 72; 73; 78; 84]. Although the controllability of T2I models has improved over the years [35; 38; 48; 82], discovering *editable semantics* remains an ongoing challenge [10; 69]. For example, as seen in the edited images with green borders in Figure 1, some prompts (e.g., “hot pink” and “tie-dye pattern” in the dress domain) can be successfully edited by the underlying T2I model (e.g., InstructPix2Pix [6], FLUX.1



Fig. 1: Motivating Examples. T2I editing methods like InstructPix2Pix [6] and FLUX.1 Kontext [3] can apply some semantics, but fail on others. This underscores the need for an automated approach to uncover the full range of editable semantics in generative models.

Kontext [3], etc.). However, other prompts cannot be edited by the model (e.g., “tulip hem,” “lower waistline,” etc. in the dress domain), as indicated by the edited images with red borders in Figure 1. Thus, determining which aspects of an image are editable is a time-consuming process, particularly when it comes to crafting effective prompts, setting guidance scales, choosing suitable images, and tuning other model-specific parameters.

Finding a large set of editable semantics that are tailored to a specific domain is also an ongoing effort in T2I editing [41]. Multimodal vision-language models can make this process easier by automating the generation [2; 55] of relevant semantics for a domain based on the user’s prompt. However, using language models comes with its own set of limitations [1; 8; 30; 81], such as hallucinations or dependence on the original prompt [42; 45], which can potentially influence the outputs of the downstream T2I model. Furthermore, it is also difficult for users to understand the full range of capabilities (e.g., creative image generation [22; 70; 77]) offered by T2I editing models firsthand. Previous works have focused on improving the diversity of image generation models (e.g., GANs, diffusion, etc.) to reduce the production of similar outputs [12; 66; 74; 83], while relatively few have explored the same problem in the context of image editing [62; 64]. Motivated by these challenges, we propose RankT2I as a potential solution to this problem. Our key contributions are as follows:

- We introduce RankT2I, a training-free, model-agnostic framework for automatically discovering and ranking editable semantics in T2I models. This allows users to identify and apply meaningful edits to their image using the model of their choice without requiring significant expertise.
- We formulate semantic edit discovery as a set selection problem and propose a submodular objective that jointly optimizes semantic relevance, model editability, and visual diversity by leveraging text prompts with unedited and edited images.
- We apply our approach across various domains (e.g., fashion, landscapes, artwork, and faces) and demonstrate through quantitative metrics that RankT2I

consistently outperforms existing diffusion- and FLUX-based semantic discovery methods by finding more diverse and interpretable image edits in significantly less time.

2 Related Work

2.1 Text-to-Image (T2I) Editing Models

Text-to-image editing models have seen substantial progress in recent years, with diffusion- and FLUX-based models emerging as two dominant approaches. Diffusion-based models work by adding Gaussian noise over several time steps and then training a neural network to recreate the original image by reversing this process [26; 58; 65]. Several diffusion-based T2I models have improved image editing by using implicit masking [5], correcting edits by shifting the denoising process [15], using a stochastic differential equation to iteratively remove the noise from the generated image [47], training a T2I model based on synthetic images and text prompts [6], using null-text optimization [49], and incorporating conditioning layouts in T2I models [82].

Compared to diffusion models, rectified flow models can also be used to generate images by applying an ordinary differential equation (ODE) to connect two distributions [18; 40; 43]. A recent example of this is FLUX.1 Kontext, which depends on a velocity prediction objective to generate images [3]. Other examples of T2I rectified-flow-based models directly find mappings between original and edit text prompt distributions [33], are used for inversion tasks [59], and apply both global and localized editing methods within FLUX [13]. Despite these advances, semantic discovery remains challenging, as some text prompts do not translate into editable visual changes.

2.2 Finding Interpretable Directions in T2I Models

Identifying interpretable directions for image editing is a difficult task and was first explored in generative adversarial networks (GANs) [20; 24; 27; 61; 76; 79]. Recently, a few works have explored applying semantic discovery to diffusion and FLUX-based T2I editing models. NoiseCLR [14] aims to find editable directions in diffusion models by training a contrastive learning objective. However, interpreting the directions identified by NoiseCLR is challenging because the resulting edits depend on the editing scale and the chosen denoising timesteps. Determining the appropriate scale and timestep settings for a given attribute can be difficult, particularly for users with limited T2I editing experience. Additionally, SliderSpace [19] applies principal component analysis (PCA) to a set of generated images, extracts their CLIP features, and trains a slider for each principal component to identify interpretable directions. However, this process is inefficient; the discovered semantics require labeling post-discovery, and the current editing workflow only supports images generated by the text prompt. The quality of the resulting sliders also depends on the number of images generated for the target domain. For example, larger samples can lead to more robust

directions but consequently slow down the semantic discovery process. Furthermore, selecting an appropriate editing scale is challenging, as large-scale values can potentially alter the identity of the original subject or distort the image.

Additionally, [85] finds interpretable directions in diffusion models by extending the framework proposed in [71] for GANs, but requires training. Furthermore, [21] uses PCA and manipulates vectors in an h-space to identify semantic directions in diffusion models; however, this process relies on classifiers to label the edits, which may limit the directions that are discovered as each direction corresponds to a classifier-defined label. Compared to these existing methods, RankT2I is training-free and discovers several diverse, interpretable semantics across multiple domains for both diffusion and FLUX-based models.

3 Methodology

To identify a set of relevant, editable, and diverse semantics for image editing, we follow a three-step process: (1) generate a list of candidate semantics, (2) create a set of edited images, and (3) use a submodular objective function to discover interpretable directions. Figure 2 illustrates this process.

3.1 Collection of Semantics

Given a user prompt tailored to a specific domain, we use a multimodal vision-language model [63] to generate a diverse set of candidate semantics for image editing. This produces keywords and phrases capturing a wide range of attributes, styles, and objects without manual specification. We denote this set of candidate semantics as $\mathcal{T} = \{t_1, t_2, \dots, t_M\}$. To evaluate each semantic t_i , we apply it to N original (unedited) images sampled from the domain.

Using a pre-trained T2I diffusion or FLUX-based model [3; 5; 6], we generate a set of edited images E_i (where each edited image corresponds to a semantic $t_i \in \mathcal{T}$ and where n_i is the total number of edited images generated for t_i):

$$E_i = \{x_{i,1}^{\text{edited}}, x_{i,2}^{\text{edited}}, \dots, x_{i,n_i}^{\text{edited}}\}$$

To determine how well a semantic can be edited within the target domain, we use the results from all of the edited images associated with that semantic.

3.2 RankT2I

Given the candidate semantic set $\mathcal{T} = \{t_1, t_2, \dots, t_M\}$, our goal is to select a subset S that is representative and diverse. Presenting all candidates to the user is impractical when M is large (e.g., 500–1000), and many semantics may be redundant or correspond to minor visual variations. We therefore formulate semantic discovery as a subset selection problem: selecting $S \subseteq \mathcal{T}$ with size $K \ll M$ that prioritizes semantics that are relevant to the domain, reliably editable, and visually diverse. The resulting subset provides a concise summary

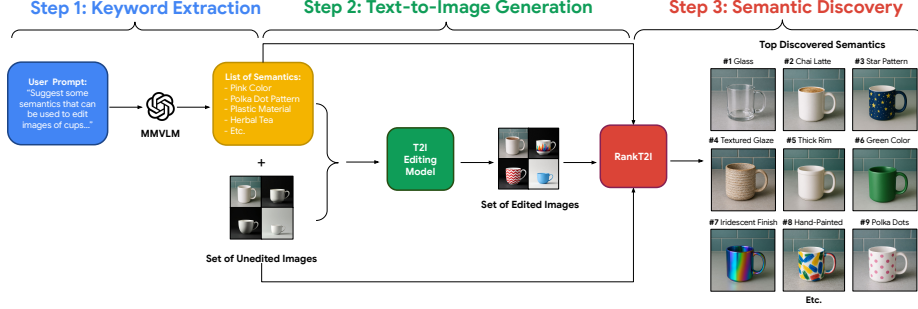


Fig. 2: An Overview of RankT2I. First, we prompt a multimodal vision-language model (MMVLM) [63] to obtain a list of potentially editable semantics for a domain. Then, we feed these semantics and a set of unedited images into a T2I model [3; 5; 6], to get a set of edited images. Lastly, we input the attributes list, unedited images, and edited images into our submodular framework, “RankT2I,” which suggests relevant, editable, and diverse semantics.

of the editing space, offering users representative and high-quality edit options without overwhelming them with redundant suggestions. The subset selection algorithm itself operates entirely over $\mathcal{T}$, where each candidate corresponds to one semantic attribute.

We propose a submodular objective function that selects a *relevant*, *editable*, and *diverse* subset ($S \subseteq \mathcal{T}$) of semantics for image editing, which is defined as follows:

$$F(S) = \sum_{t_i \in S} \left[\underbrace{\alpha F_{r_i}}_{\text{(Relevance)}} + \underbrace{\beta F_{e_i}}_{\text{(Editability)}} \right] + \lambda \underbrace{\sum_{t_j \in \mathcal{T}} \max_{t_i \in S} s(t_i, t_j)}_{\text{(Diversity via coverage)}}. \quad (1)$$

In Equation 1, α , β , and λ are weighting coefficients for relevance, editability, and diversity, respectively. Specifically, t_i represents the candidate semantic that is currently being evaluated. On the other hand, t_j represents a semantic attribute in the candidate set $\mathcal{T}$ and is used when measuring semantic coverage and diversity. The function $s(t_i, t_j)$ is the cosine similarity between the CLIP text embeddings of semantics t_i and t_j , which is used to measure the coverage of various semantics. The key components of this submodular objective function are described below.

Relevance (F_{r_i}): For a particular semantic t_i , we compute the CLIP similarity between every edited image generated using t_i and the corresponding semantic t_i . The relevance score is the average over all edited images (where j indexes the images):

$$F_{r_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \text{CLIP}(x_{i,j}^{\text{edited}}, t_i), \quad (2)$$

This measures how well the edits generated for semantic attribute t_i align with the corresponding semantic description, averaged over all available edited samples associated with that semantic.

Editability (F_{e_i}): The editability score measures how well an edit is added to the original image while preserving the identity of the subject in the original image:

$$F_{e_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\text{CLIP}(x_{i,j}^{\text{edited}}, t_i) - \text{CLIP}(x_j^{\text{orig}}, t_i) \right] \text{CLIP}(x_{i,j}^{\text{edited}}, x_j^{\text{orig}}). \quad (3)$$

The first term measures the increase in CLIP similarity to the semantic attribute t_i introduced by the edit (e.g., how much more the edited image $x_{i,j}^{\text{edited}}$ aligns with t_i compared to the original image x_j^{orig}). This gain is multiplied by $\text{CLIP}(x_{i,j}^{\text{edited}}, x_j^{\text{orig}})$, the CLIP-based image-to-image similarity between the edited and original images, to encourage edits that preserve the visual content of the source image. Averaging this quantity over all available edited samples yields the editability score F_{e_i} for semantic attribute t_i . A high F_{e_i} therefore indicates that the semantic can be introduced consistently across images while maintaining strong visual similarity to the corresponding originals.

Algorithm 1 RankT2I

```

1: Input: Candidate semantics  $\mathcal{T} = \{t_1, \dots, t_M\}$ , weights  $\alpha, \beta, \lambda$ , budget  $K$ 
2: Output: Ranked subset  $S \subseteq \mathcal{T}$ , with  $|S| = K$ 
3: Initialize  $S \leftarrow \emptyset$ 
4: Initialize  $m_j \leftarrow 0, \forall t_j \in \mathcal{T}$  ▷ Current coverage of each semantic
5: Compute  $\Phi(t_i) \leftarrow \alpha F_{r_i} + \beta F_{e_i}, \forall t_i \in \mathcal{T}$ 
6: for  $\ell = 1$  to  $K$  do
7:    $t^* \leftarrow \text{None}, \Delta^* \leftarrow -\infty$ 
8:   for all  $t_i \in \mathcal{T} \setminus S$  do
9:      $\Delta_{\text{div}}(t_i | S) \leftarrow \sum_{t_j \in \mathcal{T}} [\max\{s(t_i, t_j), m_j\} - m_j]$ 
10:     $\Delta F(t_i | S) \leftarrow \Phi(t_i) + \lambda \Delta_{\text{div}}(t_i | S)$ 
11:    if  $\Delta F(t_i | S) > \Delta^*$  then
12:       $\Delta^* \leftarrow \Delta F(t_i | S)$ 
13:       $t^* \leftarrow t_i$ 
14:    end if
15:  end for
16:   $S \leftarrow S \cup \{t^*\}$ 
17:   $m_j \leftarrow \max\{m_j, s(t^*, t_j)\}, \forall t_j \in \mathcal{T}$ 
18: end for
19: return  $S$ 

```

Diversity: The second summation term is a diversity reward, defined as a facility-location function [37] that encourages the selected set S to cover the entire semantic space $\mathcal{T}$; λ controls its contribution. To obtain the ranked list of

semantic edits, we maximize $F(S)$ under a budget (e.g., select the top K edits to present). To optimize this function, we adopt a greedy strategy that is optimal up to a known constant factor $(1 - 1/e) \approx 63\%$ [50; 67]. The greedy algorithm starts with $S = \emptyset$ and iteratively adds the candidate $t_i \notin S$ that yields the largest marginal gain (e.g., $\Delta F = F(S \cup t_i) - F(S)$). At each iteration, only semantic attributes that have not yet been selected are considered; we continue this process until $|S| = K$. Because F is submodular, each addition has a *diminishing returns* effect: once a certain semantic niche (e.g., *edits related to sleeve types in the dress domain*) is filled by one concept, other similar concepts will have lower marginal gains. Note that since the editability score can be negative when an edit reduces alignment with the target semantic, we clip each candidate’s combined score to be non-negative. This ensures our function is monotone in addition to submodular, so the greedy procedure retains its approximation guarantee.

To perform the selection, we maintain for every semantic attribute $t_j \in \mathcal{T}$ a running coverage value:

$$m_j = \max_{t_{i'} \in S} s(t_{i'}, t_j), \quad (4)$$

which is initialized to $m_j = 0$ when $S = \emptyset$. At each iteration, we evaluate the marginal gain of each remaining semantic attribute $t_i \notin S$ as

$$\Delta F(t_i | S) = \Phi(t_i) + \lambda \sum_{t_j \in \mathcal{T}} [\max\{s(t_i, t_j), m_j\} - m_j] \quad (5)$$

In Equation 5, $\Phi(t_i) = \alpha F_{r_i} + \beta F_{e_i}$ denotes the combined relevance and editability score associated with semantic attribute t_i . This computation is efficient since for each semantic attribute t_j , the diversity contribution is:

$$\lambda (\max\{s(t_i, t_j), m_j\} - m_j),$$

which is positive only when selecting t_i improves the current coverage of t_j . After selecting the semantic attribute with the largest marginal gain, we incrementally update the coverage values as:

$$m_j \leftarrow \max\{m_j, s(t_i, t_j)\}, \quad \forall t_j \in \mathcal{T}.$$

We repeat this process until the selected set satisfies $|S| = K$ and finally output the ranked list of semantics to the user. The pseudocode is outlined in Algorithm 1.

4 Experiments

We demonstrate RankT2I’s effectiveness through three core tasks: (1) uncovering a broad set of diverse, relevant, and editable semantics across multiple domains, (2) qualitatively and quantitatively assessing the diversity and interpretability of the discovered attributes against existing semantic discovery and attribute retrieval methods, and (3) performing ablation studies to reveal how

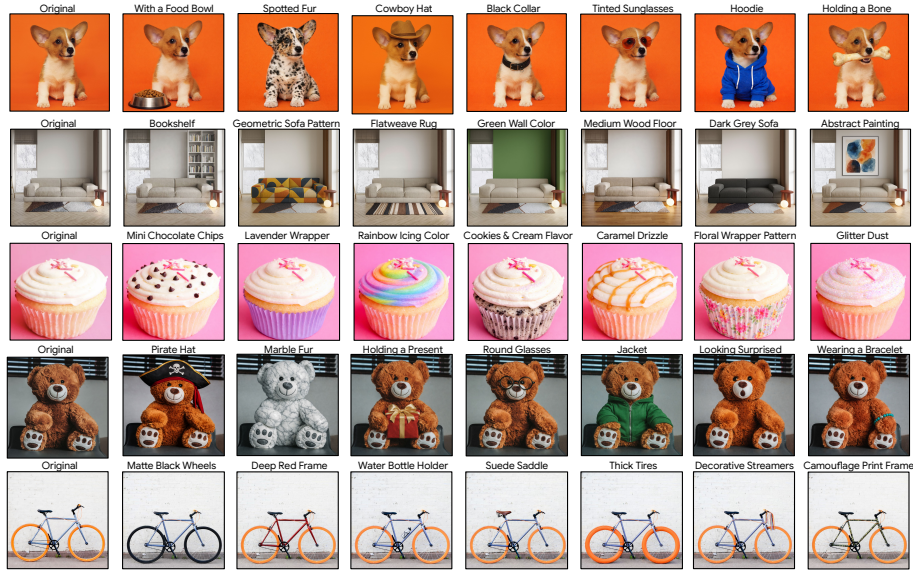


Fig. 3: Semantic Discovery for Real-Image Editing Using a Closed-Source Model. RankT2I finds both diverse and interpretable semantics in real images in many domains. GPT-Image-1.5 (see Footnote 1) was the base model that was used to generate these edits. The original images are from Unsplash (see Footnote 2).

hyperparameters and model choices impact the final rankings. In the supplementary material, we include experiments demonstrating the generalizability of these semantics across images and additional experimental details (e.g., prompts).

Experimental Setup. All experiments were performed using an NVIDIA L40 GPU, with the exception of the generated SliderSpace FLUX images in Figure 4 and the FLUX-based time comparisons in Table 3, since SliderSpace requires about 48GB VRAM to run. In these cases, an NVIDIA H100 SXM5 GPU was used. Additionally, for RankT2I, the diffusion-based edited images were created using Ledits++ [5], and InstructPix2Pix [6], while FLUX.1 Kontext [3] was used to generate the FLUX-based edited images. For Figure 3, we showcase RankT2I’s ability to successfully find editable semantics in real images using a closed-source model. In this figure, GPT-Image-1.5¹ was used as the editing model, and the free, original images were obtained from Unsplash². Unless otherwise noted, the rest of the original images in the figures were generated using GPT-Image-1.5 (see Footnote 1), SDXL [54], and FLUX Schnell³ [34].

As for the images used in RankT2I, we used $N=5$ unedited images per domain in Figures 4 and 5; as for Figure 3, we generated edits for a single image ($N=1$)

¹ <https://developers.openai.com/api/docs/models/gpt-image-1.5>

² <https://unsplash.com/>

³ <https://huggingface.co/black-forest-labs/FLUX.1-schnell>

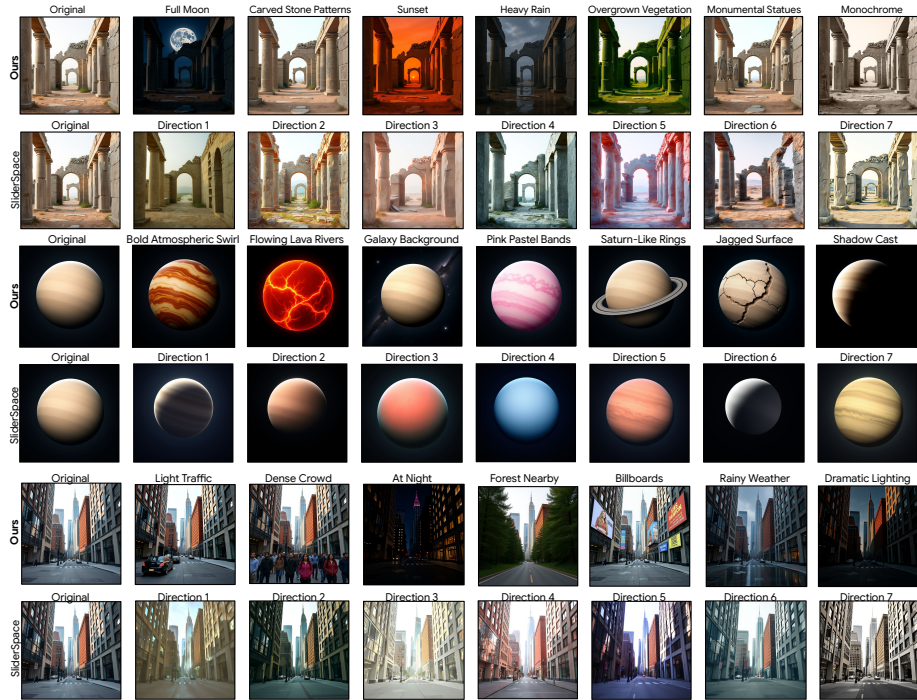


Fig. 4: Top Discovered Semantics in FLUX-Based Models. RankT2I successfully identifies several semantics in FLUX-based T2I models compared to SliderSpace across various domains, including ancient ruins, planets, and cityscapes.

to show how RankT2I can be used to find semantics specific to a particular image rather than an entire domain.

4.1 Qualitative Experiments

Figures 4 and 5 illustrate that RankT2I consistently identifies a diverse, relevant, and highly editable set of semantics across a wide range of domains, including landscapes, art, faces, etc., compared to existing diffusion-based and FLUX-based semantic discovery methods. Specifically, we compared RankT2I with other semantic discovery methods that have publicly available code (NoiseCLR and SliderSpace). Since these methods do not rank their discovered directions, we used the first-K tokens or sliders to generate the top images. For example, in the face domain in diffusion models, RankT2I suggests a wide-variety of attributes such as hair color, shirt pattern, mustache type, etc., all of which can be directly edited with the selected T2I model (Ledit++ [5] in this case). Unlike RankT2I, some directions appear visually similar across different tokens and sliders in NoiseCLR and SliderSpace. For example, in the landscapes domain, directions 3 and 4 from SliderSpace both appear to add rows to the field

of grass, while in the face domain, directions 1–3 from NoiseCLR all appear to add a green tie to the man (see Figure 5). RankT2I also identifies diverse attributes in various domains in FLUX T2I models. For instance, in the cityscapes and ancient ruins domains, RankT2I discovers a range of edits such as different times of day (e.g., “at night” and “sunset”), whereas the directions discovered by SliderSpace mainly appear to modify the color tone of the image in Figure 4. Additionally, Figure 7 shows that our method discovers semantic directions that can be used to edit various images across a domain (and are not limited to a single image).

4.2 Quantitative Experiments

To quantitatively evaluate RankT2I’s ability to discover diverse and interpretable attributes for diffusion-based and FLUX-based image editing, we compared it with NoiseCLR and SliderSpace using several benchmarks. For a fair comparison, we generated 5,000 images per method (25 directions, 200 images each) for a total of 15,000 images in the diffusion-based comparisons and 10,000 images in the FLUX-based comparisons. We assessed the diversity of the discovered attributes using the truncated CLIP entropy (TCE) [29], truncated inception entropy (TIE) [29], and pairwise image similarity using image-to-image (I2I) CLIP [25] scores. TCE measures the diversity of various semantics in CLIP’s embedding space, TIE calculates the varied visual appearance through InceptionV3 [68], and I2I finds the pairwise similarity between sets of edited images using CLIP. When interpreting TCE and TIE scores, higher values indicate images with higher diversity; for I2I, lower values indicate more visually diverse sets of images.

Furthermore, we evaluated the interpretability of the discovered semantics using the CLIP-T [56], TIFA [28], and VQAScore [39] metrics. CLIP-T finds the semantic alignment between an image and a text description through a contrastive objective. VQAScore measures text–image alignment by asking a VQA model a yes-or-no question about whether the image depicts the prompt, whereas TIFA measures alignment by generating multiple questions about specific objects, attributes, and relationships in the prompt and verifying each with a VQA model. We used the discovered semantics from the original prompts during the image generation process as the text descriptions when calculating the CLIP-T, TIFA, and VQAScore metrics for RankT2I. Since NoiseCLR and SliderSpace depend on labeling the semantics that describe the edited images after they are generated, we utilized GPT-5 [63] to generate captions describing the discovered semantics as the text prompt. As seen in Table 1, RankT2I is able to find a balance between diverse and interpretable semantics, consistently outperforming NoiseCLR and SliderSpace across these metrics.

Besides semantic discovery methods, we compared RankT2I with attribute retrieval approaches that select attributes for image editing through different strategies. For example, when editing an image, users can select semantics in several ways: randomly sampling from a list, using cosine similarity between the domain prompt and attributes to identify the most relevant ones, selecting

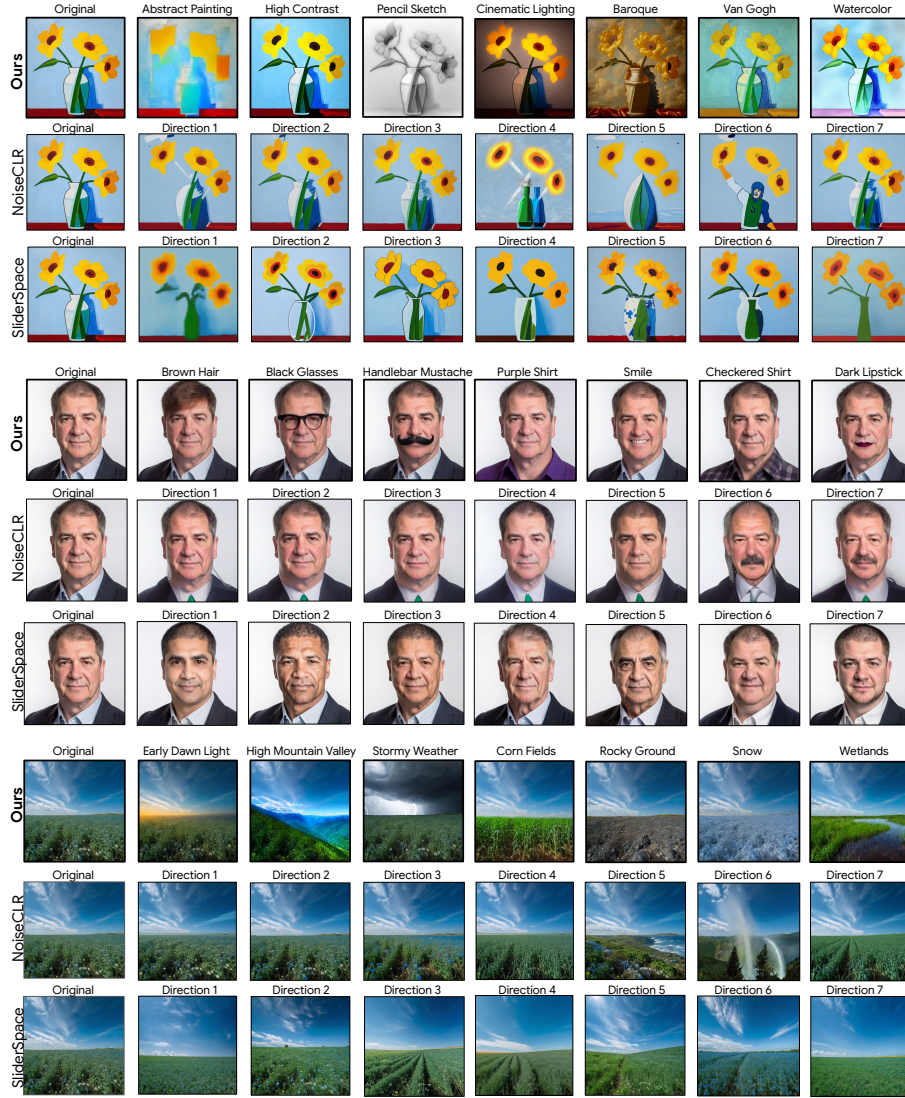


Fig. 5: Top Discovered Semantics in Diffusion-Based Models. Compared to existing diffusion-based semantic discovery methods, like NoiseCLR and SliderSpace, RankT2I successfully identifies a wide range of diverse and interpretable semantics.

the first-K attributes from a prompts list, or clustering attributes with k-means [44] and sampling one attribute from each cluster to encourage diversity. Our approach differs by examining the edited images produced by the target T2I model to uncover semantics that are specific to the model, rather than relying on predefined keywords. For both diffusion and FLUX-based models, we gener-



Fig. 6: Comparisons with Other Attribute Retrieval Methods. RankT2I successfully identifies editable, diverse, and interpretable semantics compared to other attribute retrieval methods (e.g., first-K images, random selection, k-means, etc.).

ated 5,000 images per attribute retrieval method (25 directions with 200 images each), totaling 25,000 images for the diffusion experiments and 25,000 images for the FLUX-based model experiments. Table 2 shows that RankT2I identifies both diverse and interpretable semantics across various T2I models compared to these attribute retrieval methods. Also, as seen in Figure 6, in the dress domain, selecting the first-K edited images or choosing attributes with the highest cosine similarity to a domain’s text prompt often results in redundant attributes from a single broad category (e.g., colors). Random selection can also produce redundancy (e.g., dress lengths) and may suggest attributes that the base model cannot reliably edit (e.g., “backless”). Similarly, k-means clustering may include uneditable attributes (e.g., “maxi length” or “single sleeve”). In contrast, RankT2I finds semantics that are both editable and diverse (e.g., variations in necklines, hem types, bodice designs, etc. in the dress domain).

Additionally, we compared the time it takes for our method to discover 100 semantics to NoiseCLR and SliderSpace in Table 3. Regardless of the base model (e.g., diffusion, FLUX), RankT2I takes significantly less time to find semantics for image editing than existing methods. This disparity in the total time can primarily be attributed to the training-free nature of our method.



Fig. 7: Generalizability of Semantics. The top semantics discovered by RankT2I can be applied to edit different images (Base Model: FLUX.1 Kontext).

Model	Method	$\downarrow$ I2I	$\uparrow$ TCE	$\uparrow$ TIE	$\uparrow$ CLIP-T	$\uparrow$ TIFA	$\uparrow$ VQA
SD	NoiseCLR	<u>0.8439</u>	13.8819	<u>34.2546</u>	<u>0.2378</u>	0.3402	<u>0.5484</u>
SD	SliderSpace	0.8682	<u>16.5061</u>	32.9884	0.2355	<u>0.3872</u>	0.5236
SD	Ours	0.7800	26.4701	41.1987	0.2595	0.6078	0.8177
FLUX	SliderSpace	0.9210	12.5769	35.2775	0.2542	0.3462	0.6737
FLUX	Ours	0.8405	22.8102	40.7330	0.2840	0.7159	0.9206

Table 1: Quantitative Comparisons with Semantic Discovery Methods. RankT2I identifies more diverse and interpretable attributes in both diffusion- and FLUX-based T2I models compared to existing methods.

4.3 Ablation Studies

To analyze how hyperparameters influence RankT2I, we conducted ablation studies examining the effects of λ (diversity), β (editability), and α (relevance) on the number of categories represented in the top 25 ranked attributes across multiple domains and in both diffusion and FLUX-based models. The line graphs in Figure 8 show how the number of attribute categories changes as each hyperparameter varies. When analyzing the effect of a single hyperparameter, the remaining parameters are fixed to 1. As shown in the line graphs, increasing λ leads to a higher number of represented categories, while increasing α or β reduces the number of categories. The pie charts below the line graphs in Figure 8 further illustrate how the rankings change when optimizing for only one objective in the dress domain using FLUX.1 Kontext. Ranking by relevance or editability favors a narrow set of categories, whereas diversity-based ranking improves coverage but may select less editable attributes. Overall, these results demonstrate that a balanced combination of α , β , and λ is necessary for RankT2I to recommend attributes that are simultaneously meaningful, editable, and diverse.

Model	Method	↓I2I	↑TCE	↑TIE	↑CLIP-T	↑TIFA	↑VQA
SD	First-K	0.9039	12.5140	26.4074	0.2577	0.2290	0.7012
SD	Random	0.8611	14.5577	32.4777	0.2595	0.3356	0.6972
SD	K-Means	0.8433	<u>19.1433</u>	<u>34.5400</u>	0.2635	0.4156	<u>0.7592</u>
SD	Cos. Sim.	0.9002	13.2874	27.5639	0.2764	<u>0.4846</u>	0.7459
SD	Ours	<u>0.8436</u>	19.6326	35.5105	<u>0.2688</u>	0.5021	0.8048
FLUX	First-K	0.8962	12.6479	30.7635	0.2631	0.4224	0.7028
FLUX	Random	0.9360	4.6556	26.2837	0.2450	0.4292	0.6156
FLUX	K-Means	0.8826	<u>18.1152</u>	35.1952	0.2668	0.4216	0.7936
FLUX	Cos. Sim.	0.8518	17.3142	<u>36.0747</u>	0.2920	<u>0.5840</u>	<u>0.7962</u>
FLUX	Ours	<u>0.8566</u>	20.8757	37.9963	<u>0.2799</u>	0.5995	0.8948

Table 2: Quantitative Comparisons with Attribute Retrieval Methods. RankT2I discovers more diverse and interpretable semantics than existing attribute retrieval methods.

Model	Method	Keyword	Training	Generation	Ranking	Total
SD	NoiseCLR	0	1114m 3s	3m 11s	0	1117m 14s
SD	SliderSpace	0	389m 33s	2m 15s	0	391m 48s
SD	Ours	24s	0	42m 21s	29s	43m 14s
FLUX	SliderSpace	0	287m 36s	3m 1s	0	290m 37s
FLUX	Ours	25s	0	112m 40s	1m 13s	114m 18s

Table 3: Semantic Discovery Time Comparisons. RankT2I discovers 100 semantics for text-to-image editing significantly faster than existing methods.

5 Social Implications

Like other T2I methods, RankT2I could potentially be misused for inappropriate content or deepfake generation. To help mitigate these risks, NSFW filtering during MMVLM-based keyword extraction or incorporating editing models that aim to preserve subject identity through disentangled modifications (such as Ledit++ with its implicit masking mechanism) can reduce the potential for generating harmful content. To further investigate the potential of RankT2I for mitigating deepfake generation, we conducted an experiment in the face-editing domain using FLUX.1 Kontext [3]. We generated 100 image edits and ranked them solely according to their editability scores. Our results suggest that, within this domain, semantics associated with editability scores below approximately 0.03 may be more likely to alter the identity of the original subject (see Figure 9). However, this threshold likely depends on several factors (such as the T2I editing model, domain, and set of text prompts) and may vary for other T2I editing systems. Nevertheless, the editability score provides a promising quantitative measure for identifying semantics that pose a higher risk of identity modification, potentially helping practitioners establish safeguards.

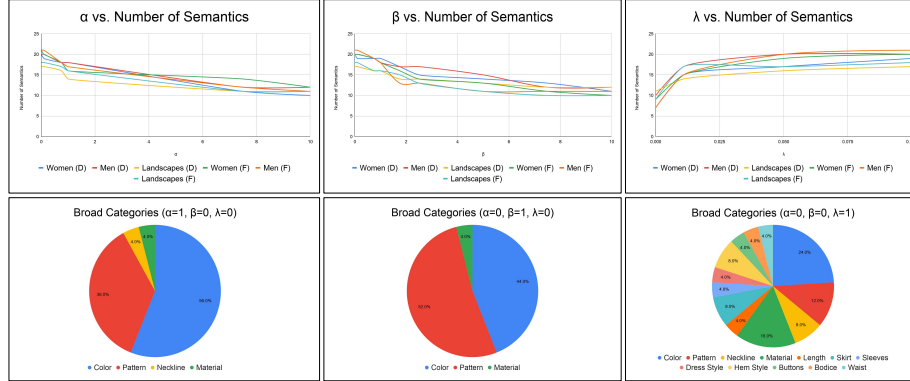


Fig. 8: Hyperparameter Ablation Studies. These charts show how changing the hyperparameters (α , β , and λ) affects the number of categories suggested by RankT2I. Note: (D) and (F) in the line graphs indicate images that were generated using diffusion- and FLUX-based T2I models.

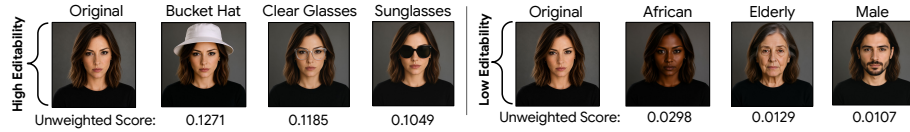


Fig. 9: Editability Score vs. Identity Preservation. As shown above, semantics with editability scores below approximately 0.03 may alter the identity of the subject.

6 Conclusion

We present RankT2I, a framework for discovering diverse, editable, and relevant semantics in T2I models. Specifically, we leverage a multimodal vision-language model to suggest a list of candidate semantics, generate edited images based on those semantics, and then apply a novel, submodular objective function to identify a set of semantics to present to users. Our approach can be applied to any T2I editing model (e.g., diffusion-based, FLUX-based, etc.) and is training-free, which significantly reduces the time needed to identify such semantics. Our experiments across multiple domains show that RankT2I uncovers a broader and more interpretable set of semantics than existing approaches, enabling applications such as stylistic image transformations as well as fine-grained object- and scene-level editing.

Acknowledgments

We thank Yusuf Dalva and Enis Simsar for their support during the early stages of this research project. This research is supported by the National Science Foundation under Grant No. 2543524.

Bibliography

- [1] Arora, D., Singh, H., et al.: Have llms advanced enough? a challenging problem solving benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7527–7543 (2023)
- [2] Bai, Y., Zhang, J., Lv, X., Zheng, L., Zhu, S., Hou, L., Dong, Y., Tang, J., Li, J.: Longwriter: Unleashing 10,000+ word generation from long context llms. In: International Conference on Learning Representations. vol. 2025, pp. 36528–36546 (2025)
- [3] Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv-2506 (2025)
- [4] Baumann, S.A., Krause, F., Neumayr, M., Stracke, N., Sevi, M., Hu, V.T., Ommer, B.: Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13231–13241 (2025)
- [5] 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. pp. 8861–8870 (2024)
- [6] 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)
- [7] Butt, M.A., Wang, K., Vazquez-Corral, J., Van De Weijer, J.: Lumictrl: Learning illuminant prompts for lighting control in personalized text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5925–5934 (2026)
- [8] Campbell, D., Rane, S., Giallanza, T., De Sabbata, N., Ghods, K., Joshi, A., Ku, A., Frankland, S.M., Griffiths, T.L., Cohen, J.D., et al.: Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems* **37**, 113436–113460 (2024)
- [9] Chen, D.Y., Tennent, H., Hsu, C.W.: Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8619–8628 (2024)
- [10] Chen, M., Zhang, Z., Zhao, C., Chen, M., Jiang, W., Jiang, T., Zhuo, J., Tang, Y., Xiao, Q., Zhang, J., et al.: Failureatlas: Mapping the failure landscape of t2i models via active exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40782–40791 (2026)

- [11] Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
- [12] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
- [13] Dalva, Y., Venkatesh, K., Yanardag, P.: Fluxspace: Disentangled semantic editing in rectified flow models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13083–13092 (2025)
- [14] Dalva, Y., Yanardag, P.: Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24209–24218 (2024)
- [15] Deutch, G., Gal, R., Garibi, D., Patashnik, O., Cohen-Or, D.: Turboedit: Text-based image editing using few-step diffusion models. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
- [16] 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)
- [17] Dunlop, C., Zheng, M., Venkatesh, K., Yanardag, P.: Personalized image editing in text-to-image diffusion models via collaborative direct preference optimization. *Advances in Neural Information Processing Systems* **38**, 167742–167771 (2026)
- [18] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
- [19] Gandikota, R., Wu, Z., Zhang, R., Bau, D., Shechtman, E., Kolkin, N.: Sliderspace: Decomposing the visual capabilities of diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15994–16003 (2025)
- [20] Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: Ganalyze: Toward visual definitions of cognitive image properties. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5744–5753 (2019)
- [21] Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., Graßhof, S., Brandt, S.S., Michaeli, T.: Discovering interpretable directions in the semantic latent space of diffusion models. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9. IEEE (2024)
- [22] Han, J., Kwon, D., Lee, G., Kim, J., Choi, J.: Enhancing creative generation on stable diffusion-based models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28609–28618 (2025)
- [23] Han, Z., Mao, C., Jiang, Z., Pan, Y., Zhang, J.: Stylebooth: Image style editing with multimodal instruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1947–1957 (2025)

- [24] Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems* **33**, 9841–9850 (2020)
- [25] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*. pp. 7514–7528 (2021)
- [26] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
- [27] Hu, H., Jiang, B., Zhou, X., Huo, X., Zhang, B.: Unsupervised discovery of disentangled interpretable directions for layer-wise gan. In: *CCF Conference on Big Data*. pp. 21–39. Springer (2022)
- [28] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 20406–20417 (2023)
- [29] Ibarrola, F., Grace, K.: Measuring diversity in co-creative image generation. *arXiv preprint arXiv:2403.13826* (2024)
- [30] Izadi, A., Banayeeanzade, M., Askari, F., Rahimiakbar, A., Vahedi, M., Hasani, H., Baghshah, M.: Visual structures help visual reasoning: Addressing the binding problem in lvlms. *Advances in Neural Information Processing Systems* **38**, 162930–162968 (2026)
- [31] Jia, Y., Cheng, A., Yuan, Y., Wang, C., Li, J., Jia, H., Zhang, S.: Designedit: Unify spatial-aware image editing via training-free inpainting with a multi-layered latent diffusion framework. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 3958–3966 (2025)
- [32] 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. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6007–6017 (2023)
- [33] Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19721–19730 (2025)
- [34] Labs, B.F.: Announcing black forest labs (Aug 2024), <https://bfl.ai/blog/24-08-01-bfl>, accessed on June 27, 2026
- [35] Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. *Advances in neural information processing systems* **32** (2019)
- [36] Li, Q., Qiu, S., Koo, K.K., Han, J., Bouyarmane, K.: Dit-vton: Diffusion transformer framework for unified multi-category virtual try-on and virtual try-all with integrated image editing. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 202–211 (2026)
- [37] Lin, H., Bilmes, J.: A class of submodular functions for document summarization. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. pp. 510–520 (2011)

- [38] Lin, K.H., Mo, S., Klingher, B., Mu, F., Zhou, B.: Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *Advances in Neural Information Processing Systems* **37**, 128911–128939 (2024)
- [39] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: *European Conference on Computer Vision*. pp. 366–384. Springer (2024)
- [40] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)
- [41] Liu, M., He, Z., Fan, Z., Wang, Q., Fung, Y.R.: Medebench: Diagnosing reliability in text-guided medical image editing. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. pp. 767–791 (2025)
- [42] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys* **55**(9), 1–35 (2023)
- [43] 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)
- [44] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. vol. 1, pp. 281–297. University of California Press (1967)
- [45] Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J.: Prompt engineering in large language models. In: *International conference on data intelligence and cognitive informatics*. pp. 387–402. Springer (2023)
- [46] Mazuz, Y., Bruner, J., Wolf, L.: Consistyle: Style diversity in training-free consistent t2i generation. *ACM Transactions on Graphics (TOG)* **44**(6), 1–16 (2025)
- [47] 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)
- [48] Meral, T.H.S., Simsar, E., Tombari, F., Yanardag, P.: Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9005–9014 (2024)
- [49] 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)
- [50] Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming* **14**(1), 265–294 (1978)
- [51] Nichol, A., 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. *arXiv preprint arXiv:2112.10741* (2021)
- [52] Orgad, H., Kavar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7053–7061 (2023)

- [53] Parihar, R., Sachidanand, V., Mani, S., Karmali, T., Venkatesh Babu, R.: Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In: European Conference on Computer Vision. pp. 469–487. Springer (2024)
- [54] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
- [55] Qu, L., Wu, S., Fei, H., Nie, L., Chua, T.S.: Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 643–654 (2023)
- [56] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastri, G., Aspell, 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)
- [57] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)
- [58] 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)
- [59] Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations. arXiv preprint arXiv:2410.10792 (2024)
- [60] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
- [61] Schwettmann, S., Hernandez, E., Bau, D., Klein, S., Andreas, J., Torralba, A.: Toward a visual concept vocabulary for gan latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6804–6812 (2021)
- [62] Simsar, E., Kocasari, U., Er, E.G., Yanardag, P.: Fantastic style channels and where to find them: A submodular framework for discovering diverse directions in gans. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4731–4740 (January 2023)
- [63] Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
- [64] Sivuk, H., Dundar, A.: Diverse semantic image editing with style codes. IEEE Transactions on Neural Networks and Learning Systems (2025)
- [65] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

- [66] Sonmezer, M., Zheng, M., Yanardag, P.: Loraverse: A submodular framework to retrieve diverse adapters for diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17879–17888 (2025)
- [67] Sviridenko, M.: A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters* **32**(1), 41–43 (2004)
- [68] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
- [69] Venkatesh, K., Dalva, Y., Lourentzou, I., Yanardag, P.: Ravel: Rare concept generation and editing via graph-driven relational guidance. *arXiv preprint arXiv:2412.09614* (2024)
- [70] Venkatesh, K., Dunlop, C., Yanardag, P.: Crea: A collaborative multi-agent framework for creative image editing and generation. *Advances in Neural Information Processing Systems* **38**, 171332–171392 (2026)
- [71] Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: International conference on machine learning. pp. 9786–9796. PMLR (2020)
- [72] Wu, C.H., De la Torre, F.: A latent space of stochastic diffusion models for zero-shot image editing and guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7378–7387 (2023)
- [73] Wu, Z., Murata, H., Takahashi, N., Wu, Q., Tsuruoka, Y.: Latentps: Image editing using latent representations in diffusion models. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 167–176 (2025)
- [74] Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2256–2265 (2021)
- [75] Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7754–7765 (2023)
- [76] Yang, H., Chai, L., Wen, Q., Zhao, S., Sun, Z., He, S.: Discovering interpretable latent space directions of gans beyond binary attributes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12177–12185 (2021)
- [77] Yang, H., Xu, K., Lu, A., Grossberg, M.D., Bai, Y., Shi, J.: Vibe spaces for creatively connecting and expressing visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21912–21921 (2026)
- [78] Ye, Y., Cai, J., Zhou, H., Li, G., Zhang, Y., Song, Z., Gao, C., Yu, J., Yang, W.: Progressive text-to-image diffusion with soft latent direction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6693–6701 (2024)
- [79] Yüksel, O.K., Simsar, E., Er, E.G., Yanardag, P.: Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In:

- Proceedings of the IEEE/CVF international conference on computer vision. pp. 14263–14272 (2021)
- [80] Zhang, C., Zhang, C., Zhang, M., Kweon, I.S., Kim, J.: Text-to-image diffusion models in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023)
 - [81] Zhang, D., Lei, J., Li, J., Wang, X., Liu, Y., Yang, Z., Li, J., Wang, W., Yang, S., Wu, J., et al.: Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9050–9061 (2025)
 - [82] 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)
 - [83] Zhang, Z., Schomaker, L.: Divergan: an efficient and effective single-stage framework for diverse text-to-image generation. *Neurocomputing* **473**, 182–198 (2022)
 - [84] 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)
 - [85] Zhang, Z., Liu, L., Lin, Z., Zhu, Y., Zhao, Z.: Unsupervised discovery of interpretable directions in h-space of pre-trained diffusion models. arXiv preprint arXiv:2310.09912 (2023)
 - [86] Zhao, H., Ma, X., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems* **37**, 3058–3093 (2024)