applied sciences

MDPI

Article

Inv-ReVersion: Enhanced Relation Inversion Based on Text-to-Image Diffusion Models

Guangzi Zhang *, Yulin Qian *, Juntao Deng and Xingquan Cai

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; dengjuntao@mail.ncut.edu.cn (J.D.); caixingquan@ncut.edu.cn (X.C.)
* Correspondence: guangzi@ncut.edu.cn (G.Z.); yulinqian@mail.ncut.edu.cn (Y.Q.)

Abstract: Diffusion models are widely recognized in image generation for their ability to produce high-quality images from text prompts. As the demand for customized models grows, various methods have emerged to capture appearance features. However, the exploration of relations between entities, another crucial aspect of images, has been limited. This study focuses on enabling models to capture and generate high-level semantic images with specific relation concepts, which is a challenging task. To this end, we introduce the Inv-ReVersion framework, which uses inverse relations text expansion to separate the feature fusion of multiple entities in images. Additionally, we employ a weighted contrastive loss to emphasize part of speech, helping the model learn more abstract relation concepts. We also propose a high-frequency suppressor to reduce the time spent on learning low-frequency details, enhancing the model's ability to generate image relations. Compared to existing baselines, our approach can more accurately generate relation concepts between entities without additional computational costs, especially in capturing abstract relation concepts.

Keywords: diffusion models; text-to-image; fine-tuning; relation inversion

Citation: Zhang, G.; Qian, Y.; Deng, J.; Cai, X. Inv-ReVersion: Enhanced Relation Inversion Based on Text-to-Image Diffusion Models. *Appl. Sci.* **2024**, *14*, 3338. https://doi.org/10.3390/app14083338

Academic Editors: Abhijit Sarkar and Lynn Abbott

Received: 8 March 2024
Revised: 8 April 2024
Accepted: 9 April 2024
Published: 15 April 2024

1. Introduction

In recent years, text-to-image diffusion models [1–4] have made significant progress. These models can generate high-quality and diverse images from text prompts written in natural language. In practical applications, there is a growing demand for personalized and customized models from users, who expect the models to generate images reflecting specific concepts. In response to this demand, recent studies such as LoRA [5], Textual Inversion [6], DreamBooth [7], and Custom Diffusion [8] have greatly enhanced the adaptability and expressiveness of diffusion models in specific downstream tasks through fine tuning [9,10].

However, while these methods excel in capturing the appearance features of images, research on exploring high-level semantics such as the relations between entities in images is scarce. In fact, the relations between entities are one of the key elements that constitute the meaning of images. Capturing these relations is not only a challenging task but also poses higher requirements for the framework: first, the model needs to understand the interactions between entities in images, which involves deep learning of complex semantics; second, the presence of relations often means that there are multiple entities in the image, and when generating images containing multiple entities, it is necessary to prevent improper fusion of features between different entities [11]; finally, given that existing diffusion model fine-tuning methods mainly focus on generating concepts with specific appearances, there is a relative lack of research in capturing relations between entities in images. In this context, ReVersion [12] pioneeringly proposed this issue and attempted to fine-tune the model to generate images with specific relations using the inversion methods.

Although ReVersion has achieved some success in generating images with specific relation concepts, it has limitations in Zero-Shot [13–15]; namely, it cannot effectively

generate uncommon entities and relations in the real world. To address this issue, we propose the Inv-ReVersion framework, which improves upon ReVersion and achieves better results (Figure 1). We designed a method of inverse relation representation in text prompts, which can effectively reduce feature fusion between entities. Based on the relation-steering loss in the ReVersion framework, we modified the structure of the contrastive loss to enrich the types of relations that can be learned. In addition, we combined digital image processing technology and designed a high-frequency suppressor from the perspective of frequency domain analysis, which helps the model focus less on image details and more on understanding the overall meaning of the images.

We implemented our framework on Stable Diffusion 1.5 [16] and conducted experiments on a dataset containing 10 different relation concepts. The experimental results show that Inv-ReVersion can significantly reduce the feature fusion problem between entities and has good reconstruction ability for complex abstract relations such as behavioral relations. We studied the contributions of each component of our framework through ablation experiments and compared our framework with existing baselines.

In summary, our main contributions are:

- We propose the inverse relation text expansion method, which can reduce feature fusion between entities in generated images.
- We propose a part-of-speech weighted control loss function, enriching the relation categories that the model can learn.
- From the perspective of the frequency domain, we designed a high-frequency suppressor, reducing the model's attention to high-frequency details, allowing it to better reconstruct the high-level semantics of images.

This study provides new insights into fine tuning pre-trained text-to-images diffusion models for high-level semantic concepts such as relations. Our proposed method advances further research in this direction.

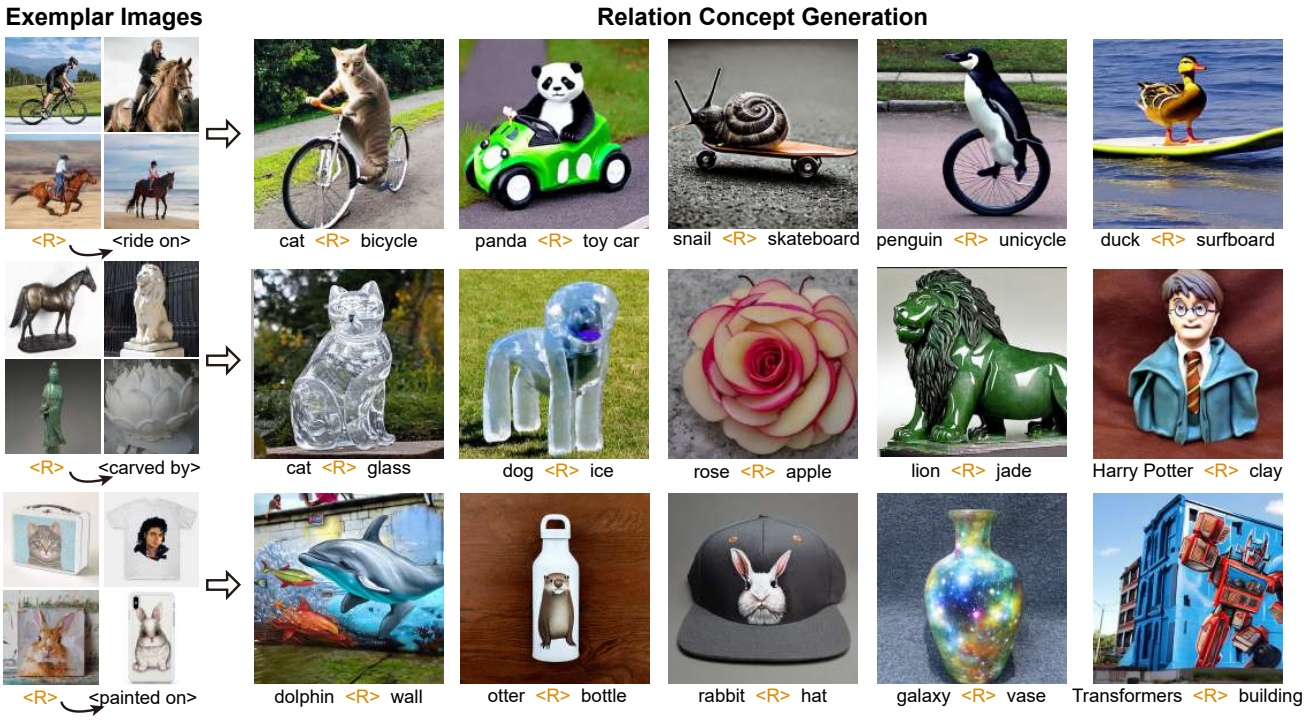

Figure 1. All images above were generated by the Inv-ReVersion framework. This task is known as Relation Inversion: it aims to capture and understand a specific relation concept from few exemplar images that share this concept and then apply it to new entities to generate images with corresponding relation features.

2. Related Work

Text-to-image diffusion models. Diffusion models [17–20] have become a mainstream method in the field of image generation [21–28] due to their ability to produce high-quality images, sparking a variety of innovative applications [29–34]. The pioneering work of this method was initially introduced by Dickstein et al. [18], who demonstrated how to use the diffusion process to simulate the gradual transition from random noise to data distribution in image generation. Subsequently, latent diffusion models (LDM) [1] proposed executing the diffusion process in the latent image space, effectively reducing the computational cost of generating high-resolution images. Text-to-image (T2I) diffusion models achieve advanced image generation performance by combining pre-trained language models, such as CLIP [35], to encode text into latent vectors. Among this series of models, Stable Diffusion, based on the LDM framework, marks the current state-of-the-art (SOTA) in this field. Our research primarily conducts experiments on the Stable Diffusion model to explore its potential in understanding and generating images with specific relations.

Diffusion-based inversion techniques mainly embed images into the model's latent space by adding noise to the images and then using the reverse denoising process of the diffusion models [29,30]. However, this process may cause significant changes to the model's latent space to adapt to the new data distribution, sometimes even leading to catastrophic forgetting in the model. To overcome this challenge, the Textual Inversion technique [6] was proposed. It aims to find a new pseudo-word in the text embedding space of the diffusion model that represents a specific concept of the image. This new word can be combined with other text prompts to form new sentences, enabling more personalized image creation. The advantage of this type of inversion method [6,7,31,36] is that it can leverage the rich semantic knowledge already learned by the text pre-training model, thus achieving model fine-tuning in an intuitive and efficient way.

ReVersion aims to extract a common relation concept $\langle R \rangle$ from several exemplar images through the Relation Inversion method. Its core idea is based on the "Preposition Prior", which sparsely activates relation prompts in the real world on a basic set of prepositions. Based on this idea, ReVersion adopts a novel relation-steer contrastive learning scheme, implemented through InfoNCE [37], to make the new word $\langle R \rangle$ representing a specific relation concept converge to the preposition space, thereby capturing interactions between entities rather than just appearance. Furthermore, it emphasizes high-level semantics between entities over low-level appearance features (such as texture or color) through relation–focal importance sampling. Extensive experiments have verified ReVersion's superiority over existing methods in a range of visual relations, demonstrating its significant advantages in visual concept embedding.

3. Method

We propose the Inv-ReVersion framework (Figure 2), aiming to enhance the model's ability to represent relations between entities by embedding vectors representing relation concepts in the model's text embedding space, through fine tuning approximately 10 sample images. This section first introduces the basic concepts of diffusion models in Section 3.1; then, in Section 3.2, we propose an inverse relation text expansion method to reduce feature fusion between entities; in Section 3.3, we introduce a part-of-speech (POS) weighted control loss function to enhance the model's learning ability for behavioral relations; finally, in Section 3.4, we approach from the frequency domain perspective, reducing the importance of detail features to enable the model to learn high-level semantic features of images more effectively.

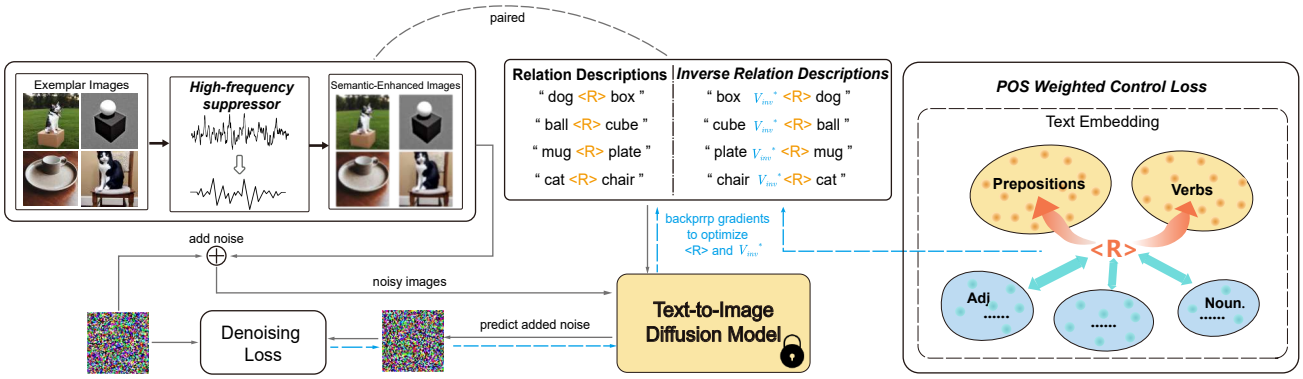

Figure 2. Inv-ReVersion framework: This framework processes exemplar images with a high-frequency suppressor to obtain semantically enhanced images. These images are then input to the model along with text prompts that have undergone inverse relation text expansion. The optimization process of the feature vector is steered by both denoising loss and part-of-speech weighted contrastive loss. In the inference phase, the optimized feature vector can be used as a new word in the text prompt to guide the model in generating images with specific relation concepts.

3.1. Preliminary

Diffusion models are a class of generative models that generate new images by gradually denoising. They start by sampling an initial image X_0 from the data distribution and gradually adding noise $\epsilon \sim N(0, I)$that follows a normal distribution. After T iterations, the noise image X_T approximates a Gaussian distribution. Subsequently, these noisy images are progressively denoised to recover the original image X_0.

Our research focuses on a specific type of diffusion model, namely LDM. Unlike traditional diffusion models that perform denoising directly in pixel space, LDM uses a pre-trained variational autoencoder (VAE) to map sample images from pixel space to latent space and then executes the diffusion process. LDM is also conditioned by a pre-trained text encoder t, such as CLIP [35] or BERT [38], to control the denoising model. The loss function of LDM is

$$L_{LDM}(\theta) := E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau_\theta(c))\|^2\right] \quad (1)$$

where ϵ represents the noise added to X_0 to generate X_t, and ϵ_θ is the noise predicted by the U-Net [39] network. The model is trained to minimize the loss between the true noise and the predicted noise. Stable diffusion is based on the LDM model, trained on a larger LAION [40] dataset, and uses CLIP as its text encoder.

3.2. Inverse Relation Text Expansion

In multi-entity image generation, the fusion of different entity features is a common and challenging problem [7,8,41,42]. For example, when attempting to generate an image of "a cat sitting back to back with a dog", it often happens that the cat exhibits dog features, or the dog has cat features, as shown in Figure 3. In the task of generating relations between entities, our research reveals a peculiar finding: the order of entities in text prompts affects the preference of the diffusion model, leading the model to focus more on the entity that appears first in the text, resulting in the later-appearing entity carrying features of the earlier-appearing entity [43–45]. Simply swapping the order of entities in the text may cause confusion in the expression of some non-interchangeable entity relations (such as active–passive relations).

To address this challenge, we propose an innovative method: inverse relation text expansion. This method introduces a special vector V_{inv}^* in the text embedding space, allowing the adjustment of the relative attention between entities without changing the semantics of the text prompt. Specifically, our text prompt is similar to "E_A $\langle R \rangle$ E_B", where E_A and E_B represent two entities in the text and $\langle R \rangle$ represents the relation vector

learned from sample images. We define an "inverse relation vector" V^*_{inv} that can invert the relation between E_A and E_B while maintaining the original meaning of the text, expanding "E_B $V^*_{inv}\langle R\rangle$ E_A" as the inverse relation text prompt, with $V^*_{inv}\langle R\rangle$ forming the inverse relation. During training, by adding inverse text prompts, we jointly optimize the inverse relation vector V^*_{inv} and the relation concept $\langle R\rangle$:

$$\langle V^*_{inv}, R\rangle = \arg\min_{\langle v\rangle,\langle r\rangle} E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau_\theta(c))\|^2\right] \tag{2}$$

Through this method, we can balance the model's attention to each entity, thereby reducing the feature fusion problem in generated images. This not only enhances the accuracy of image generation but also provides a new perspective for addressing the problem of entity feature confusion in generative models.

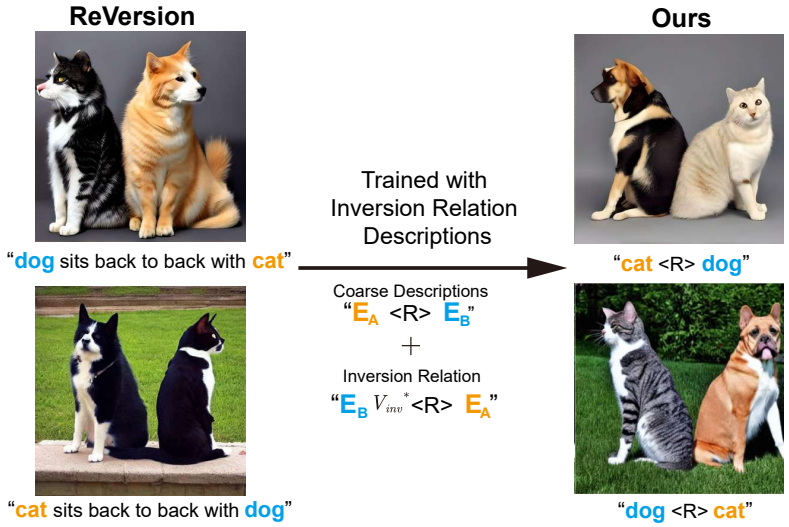

Figure 3. In the ReVersion method, as shown in the first row of the figure, when the dog appears before the cat in the text prompt, the cat on the left exhibits dog features (e.g., a longer mouth). Conversely, as shown in the second row, when the cat appears before the dog in the text prompt, the dog exhibits the patterns of the cat on the right. However, the model trained with inverse relation text expansion can effectively solve this feature fusion problem between multiple entities.

3.3. POS Weighted Control Loss

ReVersion [12] introduced "preposition prior", noting that prepositions can sparsely activate relations between entities in images. To guide the relation vector $\langle R\rangle$ into the preposition space of text embedding, ReVersion designed a relation-steering contrastive loss based on InfoNCE contrastive learning:

$$L_{\text{ReVersion}} = -\log \frac{\sum_{l=1}^{L} e^{R^T \cdot P_l^i/\gamma}}{\sum_{l=1}^{L} e^{R^T \cdot P_l^i/\gamma} + \sum_{m=1}^{M} e^{R^T \cdot N_m^i/\gamma}} \tag{3}$$

where R is the relation vector; $P_i = \{P_i^1, \ldots, P_i^L\}$ (i.e., positive samples) refers to the preposition vector randomly sampled in the ith iteration; $N_i = \{N_i^1, \ldots, N_i^M\}$ (i.e., negative samples) refers to word vectors of other POS randomly sampled from the word embedding space.

However, prepositions are mainly used to activate spatial relations in space, and they perform poorly in activating behavioral relations (such as "carved", "made of"). In contrast, verbs are often more intuitive and accurate in representing behavioral relations [46,47]. Therefore, we propose to increase the influence of verbs in contrastive loss to more effectively reinforce the embedding of relation vectors similar to behavioral relations.

Consequently, our proposed POS weighted contrastive loss is as follows:

$$L_{control} = -\left(\lambda_{verb} L_{verb} + \lambda_{prep} L_{prep}\right) \tag{4}$$

$$L_{verb} = \log \frac{\sum_{l=1}^{L} e^{R^T \cdot P_{Vi}^l / \gamma}}{\sum_{l=1}^{L} e^{R^T \cdot P_{Vi}^l / \gamma} + \sum_{m=1}^{M} e^{R^T \cdot N_i^m / \gamma}} \tag{5}$$

$$L_{prep} = \log \frac{\sum_{l=1}^{L} e^{R^T \cdot P_{Pi}^l / \gamma}}{\sum_{l=1}^{L} e^{R^T \cdot P_{Pi}^l / \gamma} + \sum_{m=1}^{M} e^{R^T \cdot N_i^m / \gamma}} \tag{6}$$

$$\lambda_{verb} + \lambda_{prep} = 1 \tag{7}$$

where L_{verb} represents the verb-steer contrastive loss, L_{prep} represents the preposition-steer contrastive loss, λ_{verb} represents the weight of the verb loss, λ_{prep} represents the weight of the preposition loss, $P_{Vi} = \left\{P_{Vi}^1, \ldots, P_{Vi}^L\right\}$ represents the verb samples, and $P_{Pi} = \left\{P_{Pi}^1, \ldots, P_{Pi}^L\right\}$ represents the preposition positive samples.

The advantage of this method is that, compared to directly replacing positive samples from prepositions with verbs, our weighting strategy can flexibly adjust the tendency of the relation vector towards spatial or behavioral relations according to the type of relation. In this way, this method will perform better in learning behavioral relation vectors than using only prepositional relations as positive samples.

3.4. High-Frequency Suppressor

Compared to appearance features, the relation concepts between entities requires the model to focus on the high-level semantic features of images. For low-level detail features, such as texture and color, the model should reduce its focus on learning [48,49]. Inspired by the field of image processing, we adopt a frequency domain analysis approach, finding that many research results indicate that the detail features of images are often associated with high-frequency features [50,51].

Therefore, our core idea is to reduce the model's attention to detail features by suppressing the appearance of high-frequency features in images, thereby better reconstructing high-dimensional semantic features (Figure 4). This approach does not affect the model's recognition of entities in the sample images. Specifically, we introduce a Gaussian filter as a low-pass filter $H(.)$ to process the input sample images X_0. At this point, the denoising loss function is defined as follows:

$$L_{denoise} = E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(H(x_t), t, \tau_\theta(c))\|^2\right], \tag{8}$$

$$H(x_t) = h * x_t \tag{9}$$

where H represents the Gaussian filter, h represents the filter kernel, and $*$ denotes the convolution operation.

In summary, the overall optimization objective of the Inv-ReVersion framework can be written as:

$$\langle V_{inv}^*, R\rangle = \arg\min_{\langle v\rangle,\langle r\rangle}(\lambda_{denoise} L_{denoise} + \lambda_{control} L_{control}) \tag{10}$$

where $\lambda_{denoise}$ is the weight of the denoising loss, and $\lambda_{contorl}$ is the weight of the part-of-speech weighted contrastive loss.

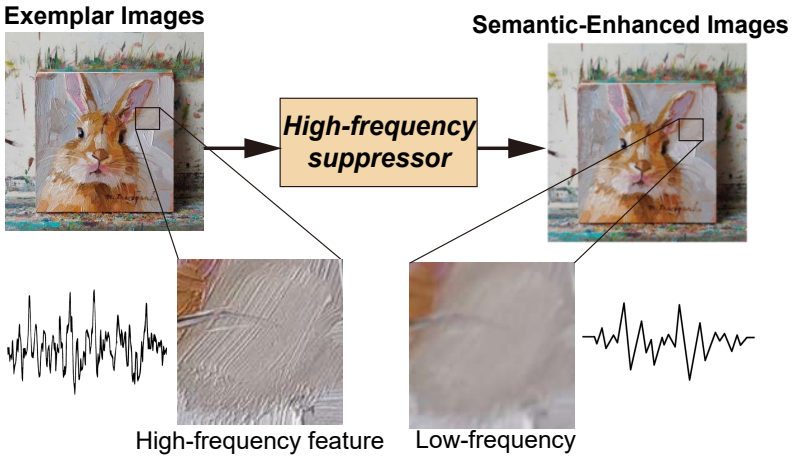

Figure 4. The left image is an exemplar image, in which high-frequency detail features such as brushstrokes are clearly visible. The right image shows the image processed by the high-frequency suppressor, where the high-frequency detail features have been suppressed. This processing reduces the model's learning of detail features, thereby improving its ability to understand the relation concepts in the exemplar images.

4. Experiments

In this section, we implemented the Inv-ReVersion framework based on latent diffusion models (LDM). Through a series of experiments, we demonstrated the advantages of our method in the task of relation inversion compared to other methods, highlighting its scientific and practical significance. The results prove that our method can effectively address the feature fusion problem between multiple entities and shows good ability in generating complex relation concepts. Finally, we discuss the limitations of our method and propose new directions for future research.

4.1. Dataset and Evaluation

Dataset. We used the ReVersion benchmark [52] for our experiments. This benchmark provides 10 different representative relation concepts at various levels of abstraction. For each relation, it provides four to ten sample images containing different entities and their text annotations.

Evaluation metrics. We invited 32 human evaluators to conduct a user study to assess the performance of our Inv-ReVersion framework in relation inversion tasks. We randomly sampled four sets of images from each of the ten concepts in the ReVersion benchmark, resulting in a total of 40 sets, each set containing images generated by four different image generation methods: Stable Diffusion 1.5, Textual Inversion, ReVersion, and Inv-ReVersion (ours). Each set also included an exemplar image describing the common relation of this set of images and the text description of each image. Evaluators were asked to rank the four results of each set according to the following criteria: (1) Entity Accuracy: Based on the text description "$E_A\langle R\rangle E_B$" of each image, evaluators need to judge whether entities A and B are correctly generated; (2) Relation Accuracy: Evaluators need to judge whether the relation between the two entities in the generated image is consistent with the coexisting relation in the exeamplar image. In this way, we obtained 80 rankings for "Entity Accuracy" and "Relation Accuracy". We then used the average user rank (AUR) [53] as a preference measure, scoring each result from 1 to 4 (lower is worse).

Baselines. We compared our method with three similar methods: Stable Diffusion 1.5, Textual Inversion, and ReVersion. Since Stable Diffusion 1.5 does not learn the relation vector in each set of exemplar images, we used natural language that best expresses the relation to replace the relation vector. For Textual Inversion and ReVersion, we used their default hyperparameters and iterated and optimized on the LDM model. All experiments

were conducted on an NVIDIA A100 PCIE 40 GB (NVIDIA Corporation, Santa Clara, CA, USA).

4.2. Qualitative Results

Figure 5 demonstrates the ability of the Inv-ReVersion framework to capture different relation concepts. Particularly in handling more complex and abstract relations, such as behavioral relations (e.g., "carved by"), the framework shows significant advantages over traditional methods that only deal with spatial relations. This capability makes Inv-ReVersion more effective in generating images containing complex interactions and dynamic scenes, thereby providing users with richer and more diverse visual content.

Furthermore, Figures 6 and 7 illustrate the notable advantages of Inv-ReVersion in addressing a challenging problem in image generation—the fusion of relations between multiple entities. In these examples, we can see that the framework can generate visually more harmonious and relationally more accurate images. Compared to other methods, Inv-ReVersion better understands and represents the complex interactions between different entities.

In summary, the Inv-ReVersion framework not only exhibits excellent performance in capturing different relation concepts but also shows significant advantages in handling the complex issue of relation fusion between multiple entities. These characteristics make the framework have broad application prospects in the field of image generation, especially in scenarios that require precise expression of complex relations.

Figure 5. Qualitative results: Using a diverse set of relation concept images as exemplars, this experiment demonstrates that our method achieves satisfactory performance in relation inversion tasks.

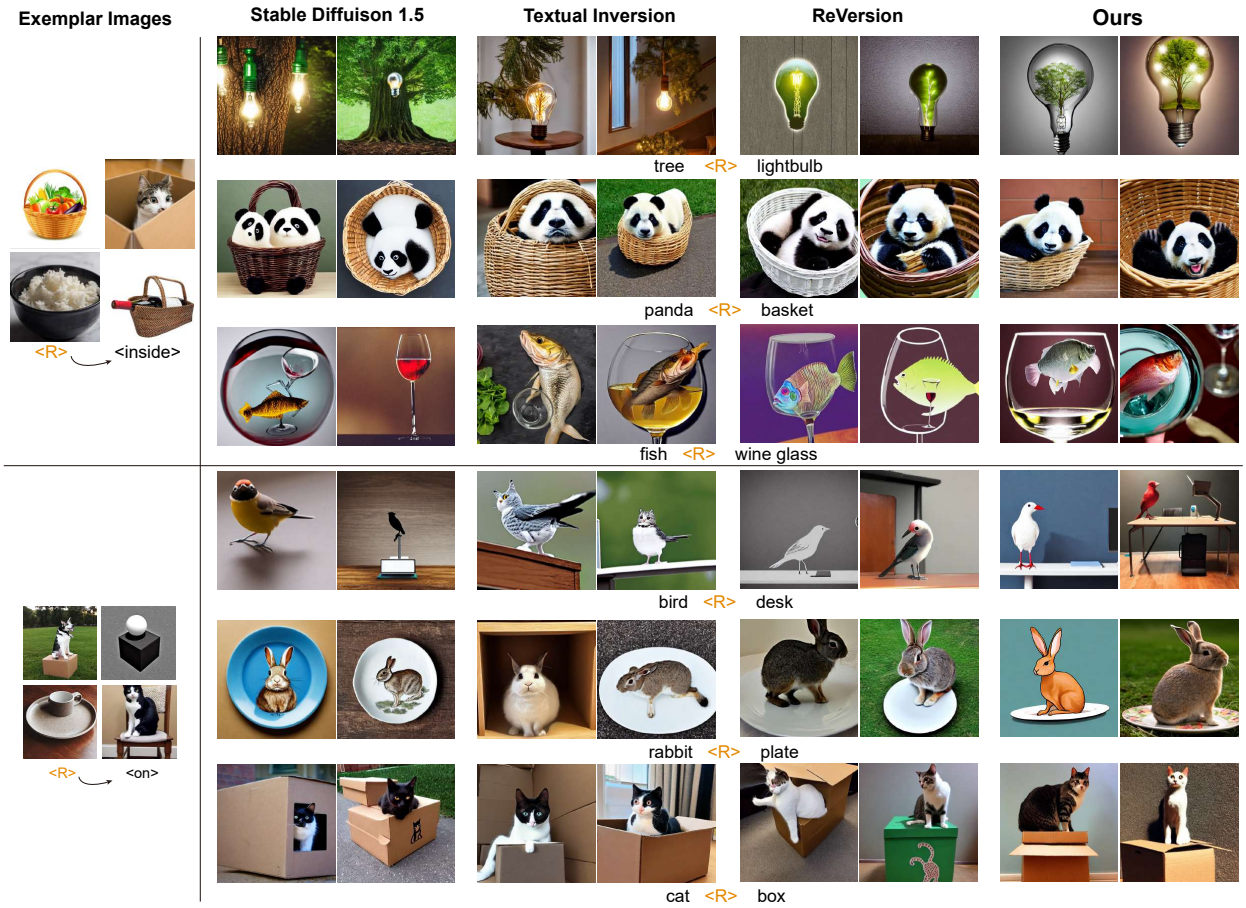

Figure 6. Quantitative comparisons: We compare the Inv-ReVersion framework with existing similar methods to validate its ability to learn common relation concepts from exemplar images. We tested the generation tasks of each relation on multiple different entities to ensure the diversity and accuracy of the tests.

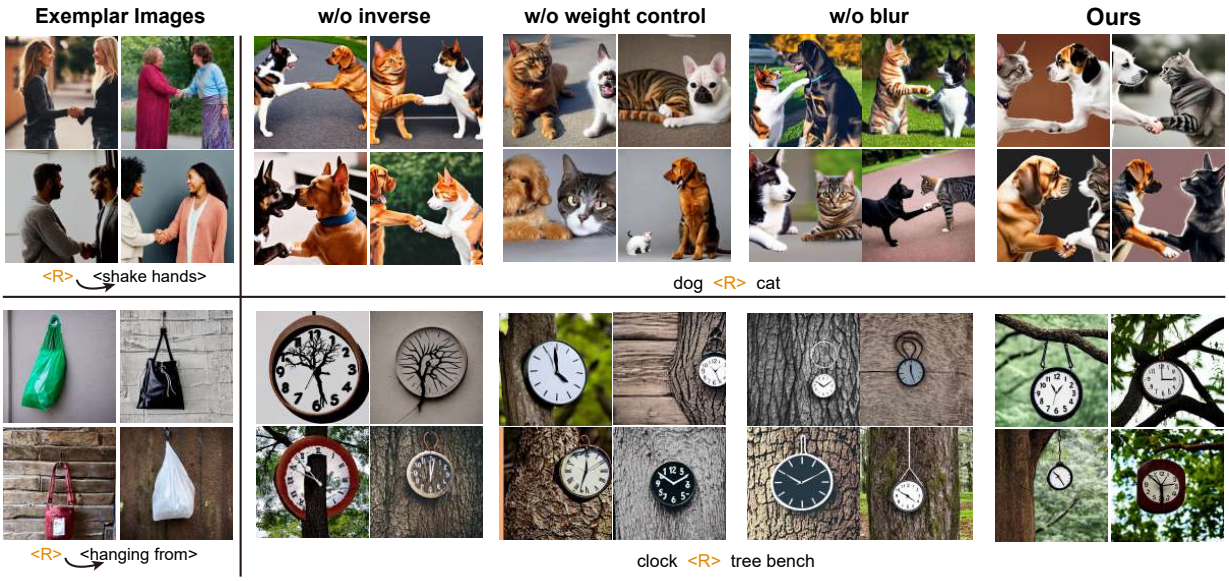

Figure 7. Ablation study: Experiments were conducted by removing each of the following components: "Inverse Relation Text Expansion", "POS Weighted Control Loss", and "High-Frequency Suppressor". The results indicate that the removal of any of these modules leads to a decrease in both "Entity Accuracy" and "Relation Accuracy" in the generated images.

4.3. Quantitative Comparisons

In our comparison experiments, we evaluated the Inv-ReVersion framework against Stable Diffusion 1.5, Textual Inversion, and ReVersion using average user ranking (AUR) as the evaluation metric. The results (Table 1) indicate that our method surpasses the other three in terms of entity accuracy and relation accuracy.

Compared to Stable Diffusion 1.5: Stable Diffusion 1.5 struggles to generate images with specific relations due to its lack of high-level semantic understanding. Even with natural language descriptions that closely represent the exemplar images, it fails to capture and generate the intended relations. Additionally, since most images in the LAION training set used for Stable Diffusion contain only one entity, the model is prone to feature fusion between entities. These issues highlight the focus of our work: enhancing text-to-image diffusion models to improve their understanding of high-level semantics and address the feature fusion problem in multi-entity images.

Compared to Textual Inversion: Textual Inversion tends to leak features of entities from exemplar images into the relation concept $\langle R \rangle$. For example, in Figure 6, the fourth row of the Textual Inversion method, the bird exhibits features of the cat's ears from the exemplar image. Moreover, Textual Inversion performs poorly in generating relations because it lacks a module specifically designed to understand high-level image semantics, leading to a focus on entity features and subpar performance in handling complex relations.

Compared to ReVersion: ReVersion, one of the best-performing methods in relation inversion tasks, can generate images with specific relations more accurately. However, it suffers from severe feature fusion between entities during the image generation process. For instance, when generating the scene "tree inside a lightbulb", (Figure 6) ReVersion merges the light of the bulb with the green of the tree, and the tree trunk takes on the shape of the filament, significantly degrading the quality of the generated image.

Overall, our Inv-ReVersion framework excels in relation inversion tasks, outperforming other methods in both entity accuracy and decoupling, as well as relation accuracy. The results of the user study (Table 1) further confirm the advantages of the Inv-ReVersion framework in these aspects, validating the effectiveness of our approach. These findings underscore the potential of Inv-ReVersion in handling complex relations and enhancing image generation quality, demonstrating its broad application prospects in future image generation tasks.

Table 1. Comparison results: average user ranking (AUR) for "Entity Accuracy" and "Relation Accuracy". We tested the user preference ranking (1–4, from worst to best) of different methods.

Concepts	Stable Diffusion		Textual Inversion		ReVersion		Ours	
	Entity	Relation	Entity	Relation	Entity	Relation	Entity	Relation
carved by	1.48	1.49	2.03	1.69	2.76	2.89	3.71	3.88
hanging	1.68	1.67	2.06	1.85	2.63	2.76	3.61	3.68
shake hand	1.83	2.13	2.04	2.04	2.76	2.81	3.32	3.04
inside	1.98	1.53	2.16	1.98	2.59	3.17	3.22	3.28
on	1.87	1.72	2.25	2.16	2.64	3.03	3.19	3.07
back2back	2.03	1.89	2.05	1.73	2.32	2.96	3.59	3.38
painted on	1.73	1.56	2.12	1.71	2.41	3.12	3.68	3.56
hug	1.93	2.14	1.97	2.01	2.56	2.58	3.52	3.24
ride on	1.69	2.03	1.74	2.14	2.92	2.78	3.63	3.03
wrapped in	1.73	1.58	2.01	1.83	2.87	3.14	3.37	3.42

4.4. Ablation Study

To verify the effectiveness of each module we proposed, we conducted the following ablative experiments: (1) Remove the "Inverse Relation Text Expansion" module; (2) Replace the "POS Weighted Control Loss" with Steer Loss from ReVersion; (3) Remove the "High-Frequency Suppressor" module. From Table 2, we can observe that the absence of these modules leads to a significant decrease in "Entity Accuracy" and "Relation Accuracy".

Without Inverse Relation Text Expansion. As described in Section 3.2, the "Inverse Relation Text Expansion" was proposed to solve the feature fusion problem between multiple entities in generated images. Based on the finding that the order of entity appearance affects model preference, we designed inverse relation text to allow entities in the text prompt to swap positions, thereby solving this problem. As can be clearly seen from the second row of Figure 7, not using inverse relation text expansion leads to the fusion of features between two entities, turning the pointers on the clock into branches.

Without POS Weighted Control Loss. Prepositions activate spatial relations well, whereas verbs have an advantage for behavioral relations. In the first row of Figure 7, the "shake hands" behavioral relation, the group without POS weighted control failed to generate images with this relation concept well. This indicates that prepositions have poor capturing ability for higher-level relations such as behavioral relations, and adding verbs can expand the relations learned in the relation inversion task.

Without High-Frequency Suppressor. Learning high-level semantic concepts like relation concepts require the model to reduce attention to low-frequency detail features. The high-frequency suppressor can solve this problem well. However, generated images without the high-frequency suppressor often carry obvious low-level features like textures, making them unable to reconstruct relation concepts well. As seen in the second row of Figure 7, the group without the high-frequency suppressor can generate detailed textures of tree bark well but failed to correctly generate the "hanging from" relation concept.

Table 2. Ablation results: Removing the inverse relation text expansion, the POS weighted control loss, or the high-frequency suppressor results in a performance decline, thereby demonstrating the importance of these three modules.

Concepts	W/O Inverse Relation		W/O Weight Control		W/O HF Suppresso		Ours	
	Entity	Relation	Entity	Relation	Entity	Relation	Entity	Relation
carved by	1.82	2.49	2.63	1.69	2.41	2.31	3.12	3.48
hanging	1.66	2.67	2.53	1.85	2.48	2.26	3.01	3.21
shake hand	1.84	2.53	2.61	1.62	2.79	2.35	2.71	3.47
inside	1.51	2.53	2.69	1.98	2.78	2.16	2.98	3.31
on	1.72	2.72	2.62	2.03	2.43	2.13	3.19	3.11
back2back	1.92	2.43	2.41	2.05	2.31	2.34	3.35	3.15
painted on	1.82	2.51	2.32	1.71	2.45	2.22	3.37	3.52
hug	1.94	2.14	2.24	2.18	2.36	2.42	3.44	3.57
ride on	1.99	2.47	2.45	1.98	2.41	2.38	3.13	3.14
wrapped in	1.76	2.38	2.41	2.06	2.51	2.26	3.31	3.26

4.5. Applications

Artistic Expansion. Traditional generative models often struggle to produce images containing multiple entities, limiting the creative freedom of artists. We propose an improved method for generating relation images, enabling generative models to produce a wider variety of artistic work. Moreover, our work involves the study of high-level semantics, which may help artificial intelligence to understand the structure, relations, and even emotions expressed in images more deeply, making artistic creation through artificial intelligence more free and interesting.

Controllable Video Generation. Video generation has been a hot research topic recently, and one challenge is the generation of relation shots. Relation shots play an important role in films, and our method can generate images with accurate relations, which can then be transformed into dynamic videos using video generation models. This breakthrough in the generation of relation shots provides a significant advancement for creating more controllable video generation workflows.

Image Editing. Changing specific attributes of an image through text prompts is a highly potential application area. Our framework can capture a variety of relation concepts, adapting to diverse image editing tasks. For example, when changing the material of an

object, as shown in the second row of Figure 1, we have endowed different objects with a variety of fresh materials using the concept of "is carved by"; in the e-commerce fashion industry, by learning the concept of "wearing", we can make models wear the product's clothes, thereby completing downstream tasks in a low-cost and efficient manner.

5. Conclusions

In this work, we conducted an in-depth exploration of the emerging topic of relation inversion by designing the Inv-ReVersion framework. To address the common issue of feature fusion in multi-entity generation, we proposed the method of inverse relation text expansion for the first time. By applying a high-frequency suppressor to sample images, we enhanced the semantic features of the images; simultaneously, through contrastive learning weighted by part of speech, we expanded the variety of relation concepts that the model can learn, enabling it to more effectively learn complex and abstract relations. Furthermore, our approach is computationally efficient, as it only updates the parameters of the text embedding, rather than all the parameters in the pre-trained model. The entire training process takes about 40 min on an Nvidia A100 PCIE 40 GB (NVIDIA Corporation, Santa Clara, CA, USA). We conducted experiments on a dataset containing multiple relation concepts, and the results verified the superiority of Inv-ReVersion in learning high-level semantics of images, surpassing existing methods. This discovery opens a new path for the research of high-level semantics in images.

Despite the high-frequency suppressor's effectiveness in identifying global relationships, it may reduce image quality by losing local details. We plan to design an additional network to restore these details and improve image generation. Additionally, while our method has shown significant effectiveness in capturing more complex relation concepts, it struggles to generate relations that are uncommon in the real world. Our next research goal is to enable relation inversion to achieve Zero-Shot Learning for generating specific relation concepts between various entities.

Author Contributions: Conceptualization, Y.Q.; Data curation, Y.Q.; Formal analysis, Y.Q.; Funding acquisition, G.Z.; Investigation, Y.Q.; Methodology, Y.Q.; Project administration, X.C.; Resources, Y.Q.; Software, Y.Q.; Supervision, G.Z.; Validation, Y.Q., J.D. and X.C.; Visualization, J.D.; Writing—original draft, Y.Q.; Writing—review and editing, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Funding Project of Humanities and Social Sciences Foundation of the Ministry of Education in China (No. 22YJAZH002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this work. These datasets can be found here: ReVersion Benchmark: https://drive.google.com/drive/folders/1FU1Ni-oDpxQCNYKo-ZLEfSGqO-j_Hw7X (accessed on 10 April 2024). Our experimental results can be found here: https://github.com/pwOliver/Inv-ReVersion-Results.git (accessed on 10 April 2024).

Acknowledgments: The authors would like to thank all who contributed to this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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, New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695.
2. 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* **2021**, arXiv:2112.10741.
3. 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. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.

4. 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* **2022**, arXiv:2211.01324.
5. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
6. Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv* **2022**, arXiv:2208.01618.
7. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22500–22510.
8. Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; Zhu, J.Y. Multi-concept customization of text-to-image diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 1931–1941.
9. Ye, H.; Zhang, J.; Liu, S.; Han, X.; Yang, W. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv* **2023**, arXiv:2308.06721.
10. Chen, J.; Zhang, A.; Shi, X.; Li, M.; Smola, A.; Yang, D. Parameter-Efficient Fine-Tuning Design Spaces. *arXiv* **2023**, arXiv:2301.01821.
11. Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; Yang, F. SVDiff: Compact parameter space for diffusion fine-tuning. *arXiv* **2023**, arXiv:2303.11305.
12. Huang, Z.; Wu, T.; Jiang, Y.; Chan, K.C.; Liu, Z. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv* **2023**, arXiv:2303.13495.
13. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv* **2022**, arXiv:2206.10789.
14. Tewel, Y.; Shalev, Y.; Schwartz, I.; Wolf, L. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 17918–17928.
15. Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; Zhu, J.Y. Zero-shot image-to-image translation. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–11.
16. Stability. Table Diffusion v1.5 Model Card. 2022. Available online: https://huggingface.co/runwayml/stable-diffusion-v1-5/ (accessed on 3 April 2024).
17. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
18. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2256–2265.
19. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, arXiv:2011.13456.
20. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
21. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
22. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 8821–8831.
23. 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* **2021**, arXiv:2108.01073.
24. Esser, P.; Rombach, R.; Blattmann, A.; Ommer, B. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3518–3532.
25. Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Trans. Graph. TOG* **2023**, *42*, 1–13. [CrossRef]
26. Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10696–10706.
27. Midjourney. Available online: https://www.midjourney.com/ (accessed on 3 January 2024).
28. OpenAI. Dall-e-3. 2023. Available online: https://openai.com/dall-e-3 (accessed on 30 March 2024).
29. Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; Zhou, J. Composer: Creative and controllable image synthesis with composable conditions. *arXiv* **2023**, arXiv:2302.09778.
30. Brack, M.; Schramowski, P.; Friedrich, F.; Hintersdorf, D.; Kersting, K. The stable artist: Steering semantics in diffusion latent space. *arXiv* **2022**, arXiv:2212.06013.
31. Iluz, S.; Vinker, Y.; Hertz, A.; Berio, D.; Cohen-Or, D.; Shamir, A. Word-as-image for semantic typography. *arXiv* **2023**, arXiv:2303.01818.
32. Poole, B.; Jain, A.; Barron, J.T.; Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv* **2022**, arXiv:2209.14988.
33. Wu, J.Z.; Ge, Y.; Wang, X.; Lei, S.W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; Shou, M.Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 7623–7633.

34. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-video: Text-to-video generation without text-video data. *arXiv* **2022**, arXiv:2209.14792.
35. 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 Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 8748–8763.
36. 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, Vancouver, BC, Canada, 18–22 June 2023; pp. 6007–6017.
37. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
39. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
40. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25278–25294.
41. Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv* **2021**, arXiv:2108.02938.
42. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
43. Hao, Y.; Chi, Z.; Dong, L.; Wei, F. Optimizing prompts for text-to-image generation. *arXiv* **2024**, arXiv:2212.09611.
44. Bar-Tal, O.; Yariv, L.; Lipman, Y.; Dekel, T. Multidiffusion: Fusing Diffusion Paths for Controlled Image Generation. 2023. Available online: https://openreview.net/forum?id=D4ajVWmgLB (accessed on 10 April 2024).
45. Liu, L.; Zhang, Z.; Ren, Y.; Huang, R.; Yin, X.; Zhao, Z. Detector Guidance for Multi-Object Text-to-Image Generation. *arXiv* **2023**, arXiv:2306.02236.
46. Insights, E. What Are Some Impressive Verbs to Use in Your Research Paper? Available online: https://www.editage.com/all-about-publication/research/impressive-verbs-to-use-in-your-research-paper (accessed on 10 April 2024).
47. Travis, C.E.; Torres Cacoullos, R. Categories and frequency: Cognition verbs in Spanish subject expression. *Languages* **2021**, *6*, 126. [CrossRef]
48. Horwath, J.P.; Zakharov, D.N.; Mégret, R.; Stach, E.A. Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images. *npj Comput. Mater.* **2020**, *6*, 108. [CrossRef]
49. Hayes, T.R.; Henderson, J.M. Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Sci. Rep.* **2021**, *11*, 18434. [CrossRef] [PubMed]
50. Xie, Z.; Zong, S.; Li, Q.; Cai, P.; Zhan, Y.; Liu, G. Interactive residual coordinate attention and contrastive learning for infrared and visible image fusion in triple frequency bands. *Sci. Rep.* **2024**, *14*, 90. [CrossRef] [PubMed]
51. Wang, F.; Eljarrat, A.; Müller, J.; Henninen, T.R.; Erni, R.; Koch, C.T. Multi-resolution convolutional neural networks for inverse problems. *Sci. Rep.* **2020**, *10*, 5730. [CrossRef] [PubMed]
52. Huang, Z. ReVersion Benchmark. 2023. Available online: https://drive.google.com/drive/folders/1FU1Ni-oDpxQCNYKo-ZLEfSGqO-j_Hw7X (accessed on 10 April 2024).
53. 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, Vancouver, BC, Canada, 18–22 June 2023; pp. 3836–3847.