Lego: Learning to Disentangle and Invert Personalized Concepts Beyond Object Appearance in Text-to-Image Diffusion Models

Saman Motamed[1], Danda Pani Paudel[1], and Luc Van Gool[1,2]

[1] INSAIT, Sofia University, Bulgaria
[2] ETH Zurich, Switzerland

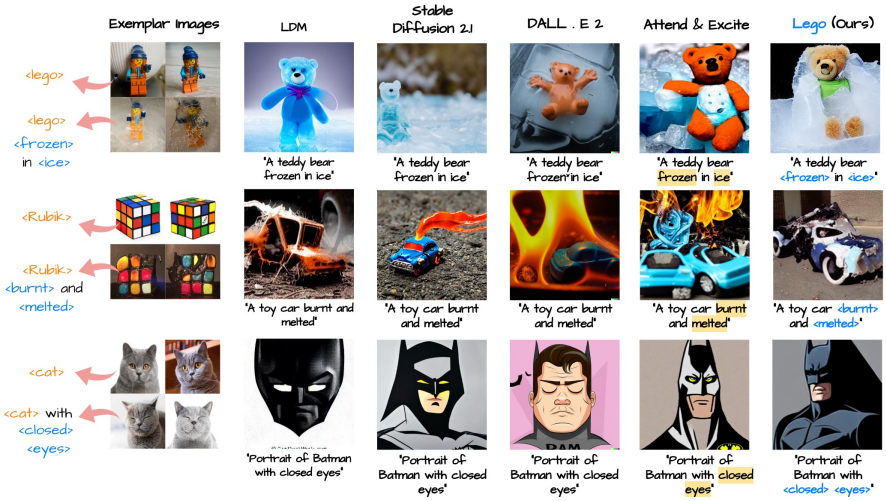

Fig. 1: We showcase Lego's ability to invert concepts of *"frozen in ice"*, *"burnt and melted"*, and *"closed eyes"* using as few as just four example images (two with and two without the concept). Our results cover text-to-image models, including LDM, Stable Diffusion 2.1, Attend and Excite, and closed-source DALL.E 2. Notably, Lego faithfully represents *intended personalized concepts*, even with a less capable backbone (LDM), while more powerful models such as DALL.E, though artistically impressive, do not consistently capture the same.

Abstract. Text-to-Image (T2I) models excel at synthesizing concepts such as nouns, appearances, and styles. To enable customized content creation based on a few example images of a concept, methods such as Textual Inversion and DreamBooth invert the desired concept and enable synthesizing it in new scenes. However, inverting personalized[3] concepts that go beyond object appearance and style (adjectives and verbs) through natural language, remains a challenge. Two key characteristics of these concepts contribute to the limitations of current inversion methods. **1)** Adjectives and verbs are entangled with nouns (subject) and can

[3] Please refer to Figure 2 for our definition of ***personalized***.

hinder appearance-based inversion methods, where the subject appearance leaks into the concept embedding and **2)** describing such concepts often extends beyond single word embeddings.
In this study, we introduce Lego, a textual inversion method designed to invert subject entangled concepts from a few example images. Lego disentangles concepts from their associated subjects using a simple yet effective **Subject Separation** step and employs a **Context Loss** that guides the inversion of single/multi-embedding concepts. In a thorough user study, Lego-generated concepts were preferred over **70%** of the time when compared to the baseline in terms of authentically generating concepts according to a reference. Additionally, visual question answering using an LLM suggested Lego-generated concepts are better aligned with the text description of the concept.

Keywords: Diffusion Models · Concept Inversion · Image Generation

1 Introduction

If you saw a Lego figurine frozen in a block of ice or a Rubik's cube melt and deform, how confident would you be in your ability to describe the fine details of the scene by using natural language descriptions alone? And even then, can text-to-image models generate images that accurately follow such text descriptions? (see Figure 1). As Mark Twain said; "Actions speak louder than words" and describing the fine details of any scene is often more difficult than showing someone / something an example of a similar scene [1, 3, 12, 24, 27, 44].

Recently, large text-to-image Diffusion models [4, 34, 42, 45, 53] have shown promising results in synthesizing high quality images. These models empower users by enabling scene synthesis through natural language descriptions. The ability to craft personalized content with these models, e.g. a scene featuring one's pet dog as Superman, has spurred a research direction aimed at enhancing the user's control for customized content creation [16, 17, 21, 23, 24, 26, 27, 31, 32, 44, 47, 51, 56]. Using a few example images of a concept, text-based inversion methods identify an embedding within the textual embedding space of a text-to-image model's text-encoder that can synthesize that specific concept. This identified embedding can then be injected into various text descriptions, allowing for the synthesis of the concept in diverse scenes. Inversion methods capable of inverting appearance-based concepts include Textual Inversion (TI) [16], DreamBooth [44], Custom-Diffusion [27] and ELITE [51]. While TI and ELITE kept the diffusion model frozen, DreamBooth and Custom-Diffusion tuned parts of the model on the example images. Taking a different approach from those works, ReVersion [24] inverts relations between subjects (e.g, under, in, etc.) from a few example images rather than concepts related to subject appearance.

With vision-language models having shown strong bias towards *nouns* / *objects* [20, 33, 37, 55] and existing inversion and personalization techniques being predominantly centered around learning appearances, relations and styles, we redirect our attention to object agnostic concepts, specifically, adjectives and

Fig. 2: A) We showcase our definition for *personalized concept* inversion. While SD 2.1 and DALL.E 2 and 3 create their version of a "frozen Lego horse in ice", we are are not only interested in synthesizing the concept, but also doing so such that it follows the example concept of the reference image (personalized) where the concept has unique characteristics (e.g. cracks and trapped bubbles in the ice). B) We visualize 4 concepts when using LDM with text description of the concept (bottom row) compared to visualizing the concepts after performing Lego inversion using reference images (visualized at the bottom of each Lego generated image) of the concept (top row).

verbs. Thus, this paper takes a comprehensive approach to examine the capabilities of text-to-image models in handling adjectives and verbs (see Figure 2) and the ability of inversion methods in learning to synthesize such concepts. We show that current inversion methods often fail to invert such concepts and our experiments suggest this challenge is due to two key characteristics inherent to these concepts. First, such concepts are entangled with a subject (noun). For instance, the concept of *melting* gives different shapes and characteristics to different subjects it is applied to and current inversion methods are not able to handle subject entangled concepts. Second, describing such concepts frequently extends beyond single word embeddings. For instance *being frozen in ice* is expressed using multiple word embeddings whereas appearance based concepts can have a single word embedding (e.g., some toy, a pet, etc.). We introduce **Lego**, a textual inversion method that augments the TI framework [16] with two additional components; a ***Subject Separation*** step that disentangles a concept from its associated subject by recovering an explicit embedding for the subject and a contrastive ***Context Loss*** that helps guide multiple embeddings in the textual embedding space, increasing editability and accuracy of the learned embeddings. We show Lego's capabilities to invert various concepts with comparisons to state of the art T2I and inversion methods, and demonstrate that Lego is a reliable and stand-alone inversion method for personalized concepts applicable to any text-conditioned diffusion model.

Our major contributions are summarized below:

- We study a new problem, ***Personalized Concept Inversion*** of adjectives and verbs. We show text-guided image synthesis models and current text based inversion and personalization methods are unable to effectively synthesize such concepts according to a given image of the concept.
- We propose two modifications to TI; **Subject Separation** and **Context Loss** that allow our method to disentangle concepts from subjects and guide

the concept's embeddings in the textual embedding space. These modifications together lead to faithful inversion of concepts.

Table 1: A characteristics overview of some recent T2I personalization methods (in order from left to right; Lego, Textual Inversion, ReVersion, Custom-Diffusion, ELITE, ControlNet and Attend and Excite), with a reference to sample images of each method when used for inverting/synthesizing adjective and verb concepts.

	Lego	TI [16]	RV [24]	C-Diff [27]	ELITE [51]	CNet [56]	A+E [7]
Frozen Model	✓	✓	✓	✗	✓	✓	✓
Text Embedding Optimization	✓	✓	✓	✓	✗	✗	✗
Subject/Concept Disentangling	✓	✗	✗	✗	✗	✗	✗
Multi-Embedding Steering	✓	✗	✗	✗	✗	✗	✗
Sample Results	Figs. 1, 7, 11	Fig. 9 - Sec. 15		Fig. 6	Fig. 15		Figs. 1, 7

2 Related Work

Diffusion Models [19, 22, 46, 48, 49] have become the go-to generative model over their counterparts [5, 6, 11, 15, 18, 25] with their superior synthesis quality and more stable training. Recently, text-to-image (T2I) diffusion models [35, 39, 40, 43, 45] have shown promise in enabling an intuitive interface for users to control image generation, using natural language descriptions. However, gaining granular control and customized content generation has proven difficult with natural language descriptions alone [14, 29, 30, 52]. Addressing this difficulty has started a line of research for inverting desired concepts in these large models and better tuning them for customized content creation. In Table 1 we describe the characteristics of the most relevant personalization models and give further detail on some of the recent works that attempt to solve this problem below.

Textual Inversion. Given a T2I model, Textual Inversion [16] is tasked with finding a pseudo-word's embedding that can represent a subject, given as few as 3-4 exemplar images of that subject. Without fine-tuning any part of the network, TI searches the textual embedding space of the diffusion model's vision-language encoder (BERT [9], CLIP [38], etc.) to find an embedding that can synthesize the given object in the reference images. In Figures 3 and 9 and Supplementary Section 15), we show that using multiple images of a concept and performing TI is not enough for inverting concepts. TI uses a single embedding that is forced to not only learn to represent the concept, but also the subject appearance, leading to appearance leakage.

DreamBooth + Custom Diffusion. With similar objective as TI, DreamBooth [44] and Custom-Diff [27] not only perform optimization in the textual embedding space, but also train parts of the T2I diffusion network with the exemplar images in order to achieve a better representation of the given concept.

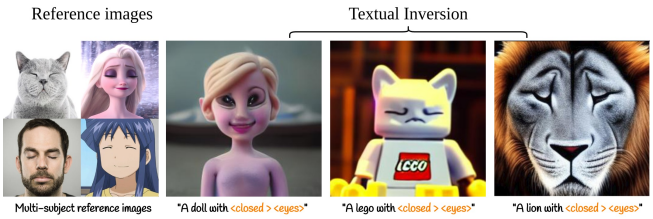

Fig. 3: Textual Inversion is not able to learn the concept of "closed eyes" from multiple subjects without the appearance of the sample subjects leaking into the concept embedding.

DreamBooth and Custom-Diff achieve better concept representation compared to TI. Custom-Diff is similar to DreamBooth with a few differences; 1) allowing multiple concepts being learned simultaneously, 2) light weight tuning by only updating the cross-attention parameters of the diffusion network and 3) a regularization step to stop language drift during tuning. Section 5.1 shows how these models fail to invert subject entangled concepts.

Attend-and-Excite. Feeding long text descriptions to T2I models often leads to catastrophic forgetting; where some words in the sentence fail to appear in the generated image. Attend and Excite [7] enforces the cross-attention units of the diffusion network to be activated for the user-selected tokens, encouraging the model to generate all subjects described in the text prompt. In Figures 1 and 7, we show that selecting the concept tokens and using the Attend and Excite method is not enough to generate the concepts.

ReVersion. Relation Inversion [24] is the only work besides ours that focuses on inverting non-appearance concepts, namely relations between subjects. ReVersion uses a contrastive loss based on InfoNCE [36] that steers the relation embedding towards the embeddings of prepositions (preposition prior), with the observation that in natural language, prepositions express the relation between subjects. ReVersion also uses natural language descriptions of the subjects in the exemplar images to separate the subjects from the relation embedding. Relations control the positioning of subjects with respect to one another and hence, the subjects stand alone. This facilitates the use of natural language descriptions alone for separating the relation embedding from subject appearance. In contrast, our focus lies in studying concepts entangled with the subject. Our ablation study (Section 5.5) shows that ReVersion's framework fails to invert subject-entangled concepts.

3 The Concept Inversion Problem

The goal of Lego is to learn embeddings $\mathcal{CPT} = \{<cpt_1>, ..., <cpt_n>\}$ that represent a concept $\mathbf{C}$, from a few exemplar images. Let $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ be a small set of such exemplar images involving a common subject $\mathbf{S}_e$. Lego requires a clear separation between images of a subset with concepts, say $\mathcal{I}_C$, and the

same without, say $\mathcal{I}_{\overline{C}}$, such that $\mathcal{I} = \mathcal{I}_C \cup \mathcal{I}_{\overline{C}}$. For the example shown in Figure 4 (right), the caption of all the images in $\mathcal{I}_{\overline{C}}$ is "photo of a Rubik's cube", whereas the same in $\mathcal{I}_C$ is "photo of a Rubik's cube that is melted", where the exemplar subject $\mathbf{S}_e$ is "Rubik's cube" and the general concept of interest $\mathbf{C}$ is "melted". In this setting, we wish to learn the embeddings $\mathcal{CPT}$ corresponding to the concept $\mathbf{C}$ such that the concept can be transferred to any novel target subject $\mathbf{S}_t$– which is "a teddy bear" in the very same example.

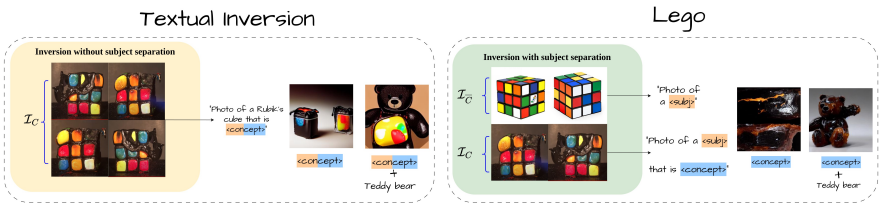

Fig. 4: Right figure is an overview of Lego's objective and the Subject Separation step. Learning an explicit embedding to represents the subject (Rubik's cube) allows the concept ("melted") embedding to dissociate from the subject's appearance features, as visualized by <concept> embedding (highlighted in blue). The left figure depicts the framework that uses concept only images (same setting as TI, DreamBooth, etc.). In this setting, the subject's features leak into the <concept> embedding, (highlighted in orange and blue), as shown by the concept's visualization which evinces both melting effects and Rubik's cube features.

Our experiments show the inherent difficulty in synthesizing accurate representations of verb and adjective concepts using T2I models with natural language guidance. The ability to learn embeddings capable of representing such concepts using as few as four exemplar images allows the user to have greater control over T2I models. In the context of T2I models, adjective and verb concept inversion, where the concept is entangled with subjects, has not been previously explored. We demonstrate that such entanglement poses challenges for inversion methods that have previously aimed at subject/concept separation, as observed in other concept types like relations [24].

4 Learning Concepts Beyond Appearance

4.1 Preliminaries

T2I Diffusion Models. Diffusion models [10, 22, 46] are a class of generative models that learn to generate novel scenes by learning to gradually denoise samples from the Gaussian prior x_T (trained by adding noise $\epsilon \sim \mathcal{N}(0, 1)$ to x_0) back to the image x_0. In this work, we study T2I models, namely the Latent Diffusion Model (LDM) [42]. Instead of directly adding noise and learning to denoise images, LDM operates on a pretrained autoencoder's projection of images in some

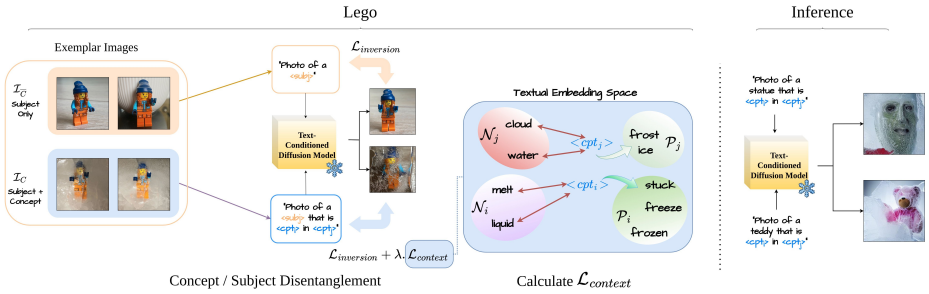

Fig. 5: An overview of Lego's framework. From left to right, during embedding optimization, Lego dedicates an embedding $< subj >$ for inverting the subject $\mathbf{S}_e$ in the exemplar images $\mathcal{I}_C$ and $\mathcal{I}_{\overline{C}}$. This stops appearance leakage to the concept embeddings. Each concept embedding ($< cpt_{i/j} >$) is separately steered towards user defined words ($\mathcal{P}_{i/j}$) that correspond to the embedding's semantic word and away from antonyms of those words ($\mathcal{N}_{i/j}$). After the inversion, the learned embeddings can together be applied to different target subjects $\mathbf{S}_t$ ("Statue" and "Teddy bear") to manifest the concept in new scenes.

latent space. LDM enables T2I generation by conditioning the denoising network $\epsilon_\theta(.)$ on the encoding of text descriptions c using a text encoder $\tau_\theta(.)$ such as CLIP [38] or BERT [9]. To sample images using a trained latent diffusion model $\epsilon_\theta(.)$, we iteratively denoise a noise latent x_t for t steps, using the predicted noise $\epsilon_\theta(\mathrm{x}_t, t, \tau_\theta(c))$ to get x_0. x_0 is then mapped to the image space using some pre-trained decoder. The LDM loss is: $\mathcal{L}_{LDM}(\theta) := \mathbb{E}_{t,\mathrm{x}_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathrm{x}_t, t, \tau_\theta(c))\|^2]$.

Inversion in the Textual Embedding Space. Current textual inversion methods focus on either appearance inversion of a subject [16, 27, 44, 50, 51] or inverting a relation between subjects [24]. Given a few exemplar images of a subject or relation, the aim is to find a text embedding $<emb^*>$ in the output space of $\epsilon_\theta(.)$, such that injecting $<emb^*>$ in any encoded text $\tau_\theta(c)$ allows the reconstruction of that concept in a new context defined by description c. The embedding inversion loss is defined as:

$$\mathcal{L}_{\text{inversion}} = \mathbb{E}_{t,\mathrm{x}_0,\epsilon}[\|\epsilon - \epsilon_\theta(\mathrm{x}_t, t, \tau_\theta(c))\|^2], \tag{1}$$

such that: $<emb^*> = \underset{<emb>}{\arg\min}(\mathcal{L}_{\text{inversion}})$, where $<emb>$ is the concept embedding being optimized, and is fed into the pretrained T2I model as part of the text description c.

4.2 Method Overview

In this section, we present **Lego**, an inversion method for extracting personalized concepts (adjectives and verbs) from exemplar images. Lego augments Textual Inversion with two modifications; ***Subject Separation*** and ***Context Loss***. Lego aims to invert concepts that move beyond objects' relations, styles, and appearances. These concepts are considered to be entangled with the subject

of the exemplar images. For instance, The concept of "melting" is not standalone and changes the features of the subject it is applied to, whereas in subject inversion (TI [16], Custom-Diff [27], etc.), the subject's appearance does not change and stands alone. Similarly, in relation inversion (ReVersion [24]), the relation defines how different subjects interact and does not change the subjects' appearance features.

4.3 Subject Separation

Appearance inversion focuses on low-level features in order to learn a single style or subject appearance from a few sample images. For such subject-centric inversions, a pixel level reconstruction loss (Equation 1) is often sufficient to find an embedding that represents the subject. Inverting relations between subjects (ReVersion) is a higher-level concept that requires more than a pixel level loss which ReVersion addresses using a preposition prior (Section 2). Relations between subjects however, are not entangled with the subjects' appearance, hence ReVersion is able to detach the relation embedding from the subject embedding by steering it away from the embedding of natural language words that describe the subjects (e.g., while inverting the relation of Batman and Superman sitting back to back, the relation embedding is steered away from embeddings of "Batman" and "Superman") . Our concepts however are entangled with subjects such that the same approach as ReVersion leads to appearance leakage (see Figures 4 - left and 9) and does not allow concepts to be separated from the subjects.

In order to learn embeddings $\mathcal{CPT}$ representing the concept $\mathbf{C}$ disentangled from the exemplar subject $\mathbf{S}_e$, we dedicate an additional embedding $< subj >$ to separately represent the subject $\mathbf{S}_e$. This embedding gets optimized twice using (i) $\mathcal{I}_{\overline{C}}$ and (ii) $\mathcal{I}_C$, separately. Learning $<subj>$ from subject-only images $\mathcal{I}_{\overline{C}}$ naturally enables the concept embeddings $\mathcal{CPT}$ to not have to learn the subject's appearance, while reconstructing $\mathcal{I}_C$ using both $<subj>$ and $\mathcal{CPT}$ embeddings (see Figure 4). In other words, learning the subject embedding without concept and using its embedding to generate the subject with concept during optimization allows us to better perform the desired disentanglement. Our experiments illustrate that the proposed learning setup prevents the subject-specific features to leak into the concept embedding (see Figure 2 - B). In Figures 4 and 9 and in the Supplementary Section 15, we show that the absence of the proposed Subject Separation leads to undesired outcomes. For instance, Figure 4 (left) shows how not performing Subject Separation in learning the concept of *melting* from Rubik's cube images leads to the "$<melted>$ toy bear" to have Rubik's features and learning the concept of *closed eyes* from multiple subjects in Figure 3 leads to the appearance of those subjects leaking into new subjects. In contrast, learning an explicit embedding for the cube - Figure 4 (right) - keeps the Rubik's features disentangled from the concept embedding.

4.4 Contrastive Context Guidance

Concepts of our interest (e.g., frozen in a block of ice) often require a multi-word description. Hence in this work, we allow for learning multiple word embeddings per concept. More specifically, Lego enables learning multiple embeddings ($\mathcal{CPT} = \{<cpt_1>, ..., <cpt_n>\}$) for a single concept – such that combining all n embeddings represents the concept,– unlike, TI [16] and Custom-Diff [27] which learn concepts described by a single word embedding. Inspired by success of contrastive losses for representation learning [24,28,36,54], we employ an InfoNCE-based [36] loss to learn the embeddings $\mathcal{CPT}$ in a contrastive setting.

Learning concept embeddings in the contrastive setting however requires positive and negative embedding sets. Let $\mathcal{P}_i = \{P_{ik}\}$ and $\mathcal{N}_i = \{N_{ik}\}$ respectively be the positive and negative embedding sets corresponding the concept embedding $< cpt_i >$. Note that each embedding $< cpt_i >$ corresponds to a semantic word. We form the sets $\mathcal{P}_i$ and $\mathcal{N}_i$ by embedding the synonyms and antonyms, respectively, of the semantic word corresponding to $< cpt_i >$. Please refer to Section 11 for more detail on choosing $\mathcal{P}_i$ and $\mathcal{N}_i$. In this manner, for the general concept $\mathbf{C}$, we obtain a set of triplets $\{(< cpt_i >, \mathcal{P}_i, \mathcal{N}_i)\}_{i=1}^{n}$. This triplets' set is then used to compute our modified InfoNCE loss, referred here as $\mathcal{L}_{\text{context}}$, which is given by,

$$\mathcal{L}_{\text{context}} = -\sum_{i=1}^{n} \log \frac{\sum_{k=1}^{|\mathcal{P}_i|} e^{<cpt_i>^\intercal \cdot P_{ik}}}{\sum_{k=1}^{|\mathcal{P}_i|} e^{<cpt_i>^\intercal \cdot P_{ik}} + \sum_{k=1}^{|\mathcal{N}_i|} e^{<cpt_i>^\intercal \cdot N_{ik}}}. \tag{2}$$

During the learning process, the context loss guides each concept embedding $< cpt_i >$ individually, towards the embeddings of the respective positive words and away from the negative ones. Figure 5 shows an example where we want to capture the concept of "frozen in ice" with two embeddings $< cpt_i >$ and $< cpt_j >$, one representing "frozen" and one representing "ice". When combined together in a sentence, they express the concept $\mathbf{C}$ in the exemplar images $\mathcal{I}_C$. In Section 5.5, we provide examples of concept inversion with and without the context loss. Our objective of learning descriptive multi-word concepts, enabled by our context loss, requires the embeddings to be steered towards their corresponding semantic word, each in a different part of the text-embedding space. Furthermore, the negative sets of words allow Lego's embeddings to be steered away from words that can disturb the concept inversion. By disturbing the inversion, we refer to words that are in close distance to our concept embedding, yet associated with a concept we do not want to represent. For instance, numbers are closely embedded in the text embedding space. In order to accurately invert cardinality, say number 3, we can construct a negative set of words comprised of $\{1, 2, 4, 5\}$.

4.5 Lego

Lego uses the inversion loss in Equation 1 to learn an embedding $<subj^*>$ that represents the subject of the exemplar images, by optimizing an embedding

$<subj>$. While optimizing the concept embeddings, the weighted sum of the inversion loss and our context loss is used to obtain the concept embeddings $<\mathcal{CPT}^*>$. Below we show both the subject and concept inversion losses.

$$\begin{aligned} <subj^*> &= \arg\min_{<subj>}(\mathcal{L}_{\text{inversion}}), \\ <\mathcal{CPT}^*> &= \arg\min_{<cpt>}(\mathcal{L}_{\text{inversion}} + \lambda.\mathcal{L}_{\text{context}}). \end{aligned} \tag{3}$$

These two embeddings' recoveries, corresponding to subject and concepts, can be thought as two parallel processes, being optimized at the same time, acting on the same image generator and the set of exemplar images. Such clear separation in learning process enables us to achieve the desired disentanglement.

5 Experiments

5.1 DreamBooth and Custom Diffusion

Both DreamBooth [44] and Custom-Diff [27] pursue the same objective by tuning parts of the diffusion model, while optimizing a word embedding in order to achieve better accuracy in synthesizing personalized concepts compared to methods like Textual Inversion that keep the model frozen. Since Custom-Diff allows learning multiple embeddings at the same time, we carried out our experiments using Custom-Diff. With the same exemplar images used in Lego, we trained Custom-Diff to learn the embeddings of "Rubik's cube" and "melting" from Rubik's $\mathcal{I}_{\mathcal{C}}$ and $\mathcal{I}_{\overline{\mathcal{C}}}$ images and the concept of "cat" and "closed eyes" from the cat images. Figure 6 shows that inverting concepts while training the network does not allow for dissociation of concepts from subjects and leads to appearance leakage. We show how Custom Diffusion's learned embedding for "melting", when applied to a *toy car* and a *toy cat*, inserts the Rubik's cube features in those images and similarly, the learned concept for "closed eyes" from the cat images, inserts the cat's features into images of *Batman* and *a doll*.

Fig. 6: Custom-Diff is unable to separate the concept embedding from the subject appearance. Learning "closed eyes" from cat images and "melted" from Rubik images will carry the subject features when applied to new subjects (appearance leakage).

5.2 Lego

We tested Lego's capabilities in inverting 10 various concepts; from controlling cardinality of subjects (**3**, **4**, **5**) to concepts that deform the subject (**melting**, **crumpling**), to the subject performing an action (**closed eyes**, **walking on a rope**, **arms raised**) and change of state and appearance of the subject (**frozen in ice** and **having a smiley emoji face**). For each experiment, 4 example images were used (2 with and 2 without the concept).

5.3 Results

In Figures 1 and 7, we show Lego's results for a subset of the concepts and provide comparisons with the latest SOTA methods for T2I generation. Should certain results be selectively highlighted, it may skew the reader's perception of Lego's effectiveness. We note that Lego can learn to authentically represent the concept of the example images even where more recent models such as Stable Diffusion 2.1 and DALL.E that can generate the concept, are not able to do so faithful to the example image. Similar to TI, Lego is standalone and can be applied to any T2I model. We will be releasing Stable Diffusion support under Hugging Face's Diffusers library [13]. For more examples of Lego with various subjects, please see Figure 11 in Supplementary material.

We tested Lego's concept inversion against natural language for 10 concepts, generating 200 images per concept using Lego and LDM with language control. Three concepts specified cardinality (3, 4, or 5 subjects). Participants counted subjects in each image. Lego consistently outperformed natural language guidance in producing correct subject counts (see Table 2). For the remaining seven concepts, we performed two different studies. A Visual Question Answering (VQA) large language model (Flan-T5 XL) [39] was used to answer questions about 200 images generated using Lego, versus 200 images generated using natural language descriptions of each concept (2800 total images). Our experiments showed that VQA models performed better when asked "Yes" or "No" questions about specific concepts rather than asking more general questions such as " What is this image?". For instance, we generated images of "toy bears frozen in ice". We then prompted the VQA model with the question: " Is the toy bear frozen in ice?". Table 2 shows the number of images where the VQA model confirms the concept is in the image. Consistent with the numerical concepts, Lego performs better than natural language for all concepts. In Supplementary Section 11, we give a detailed overview of the prompts used and the generated images.

While VQA suggests Lego outperforms LDM, a more comprehensive evaluation for generative models, especially concerning general concepts, involves human preference metrics. To this end, we used Amazon Mechanical Turk [2] and for each concept, paired the 200 images generated by Lego with the other 200 generated by LDM. We asked users to select the image they think best represents the concept for each pair of images. Each question was answered by 10 users (total of 14000 answers for all 7 concepts). The majority vote determined the outcome, showing that Lego was preferred over LDM in over 70% of cases

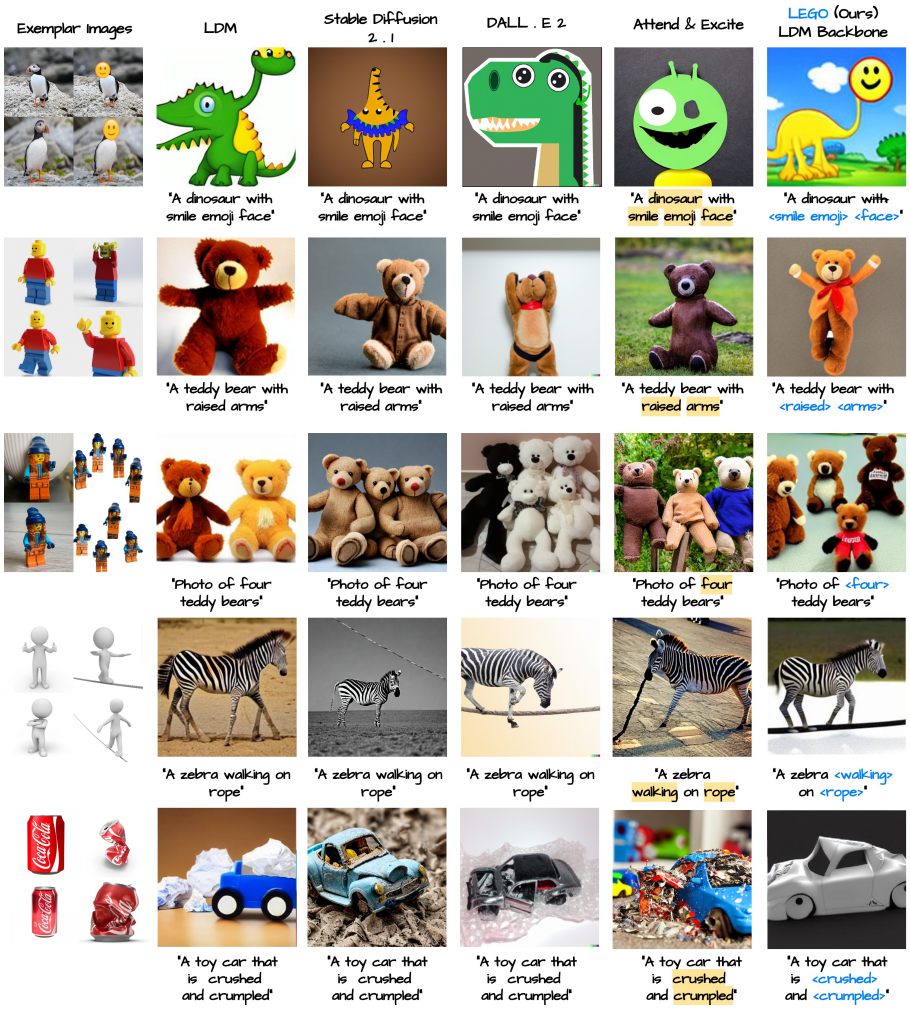

Fig. 7: Qualitative comparison of **Lego** with an LDM backbone and learned concepts from exemplar images, compared to text-guided models such as LDM, SD 2.1, DALL.E 2 and Attend & Excite (highlighted words are the given token). Zoom in for details.

Fig. 8: This figure (left) shows Lego's composition capability of combining different concepts learned from different images and (right) learning more complex multi-word embedding concepts from a single example.

(Table 2). While we use Lego to invert concepts, in Section 9 we show that the learned subject embedding is also able to perform similar to TI in inverting the subject of the reference images.

Table 2: LLM metric reports the number (out of 200) of images, where Flan-T5 XL model confirms the image contains the concept. The Human metric reports the number of images with correct subject cardinality for the numerical concepts, and the percentage of time users preferred one method's images compared to the other.

Concept	LLM ↑		Human ↑	
	Lego	LDM	Lego	LDM
3	-	-	107	85
4	-	-	63	18
5	-	-	36	18
Frozen in ice	92	55	68.5%	31.5%
Burnt and melted	136	61	77%	23%
Closed eyes	111	75	74.5%	25.5%
Smiley Emoji Face	199	151	67.5%	32.5%
Crumpled and squeezed	147	64	60.5%	39.5%
Walking on a rope	133	4	84 %	16%
Arms raised	124	49	70%	30%

5.4 Concept Composition and Complexity

Lego's learned concept embeddings can be combined, as demonstrated in Figure 8 (left). It seamlessly composes concepts like "closed eyes" from cat images and "walking on a rope" from figure images. This enables creating scenes like a lion walking on a rope with closed eyes. While we have showcased Lego's ability with mostly two-word-embedding concepts, the synthetic example in Figure 8 (right) illustrates its capability to invert more complex scenarios, learning and applying "walking on a rope with closed eyes" from a single example to the subject "lion".

5.5 Ablation Study

We studied the effect of ***Subject Separation*** and ***Context Loss*** on inverting three concepts; *"burnt and melted"*, *"closed eyes"* and *"walking on rope"*. TI requires a few images of a subject or subjects in the same style (e.g., a few Monet paintings), while ReVersion needs multi-subject images for relation inversion. For the *"burnt and melted"* concept, we used both single and multi-subject reference images to cover various settings. Our ablation study compares Lego, incorporating both Subject Separation and Context Loss, with three other combinations that remove one or both of these steps. Note that single / multi-subject experiments with neither Subject Separation nor Context Loss resemble performing

a regular Textual Inversion [16]. Performing single / multi-subject experiments by adding the Context Loss without Subject Separation resembles the ReVersion framework [24]. Figure 9 shows the ablation results for *"burnt and melted,"* highlighting the impact of Subject Separation and Context Loss for concept inversion. Complete ablation results are shown in Supplementary Section 15.

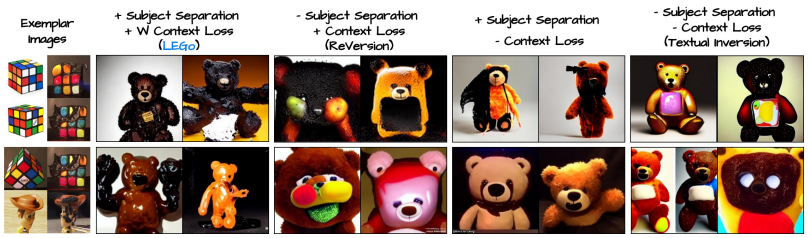

Fig. 9: This figure shows examples of our ablation study for learning the concept of *"burnt and melted"* from single subject (top row) and multi subject (bottom row) example images and transferring it to a *"Toy bear"*. We synthesized 100 images for each ablation category. You can find all 100 images for each category in the Supplementary Section 15. Please zoom in to see the details in the images.

5.6 Example Image Size + Choice of Positive / Negative Words

In Section 8 of the supplementary, we show the effect of increasing reference images on Lego's performance. Our approach to selecting positive and negative words involved choosing synonyms and antonyms of semantically meaningful words for the concept (Figures 12 and 13), demonstrating robustness without requiring adjustments. While studying word selection's subjective impact on inversion is challenging, it remains an interesting avenue to be explored

6 Conclusion

In this work, we took a first look at capabilities of T2I models in synthesizing adjective and verb concepts by reference examples. We showed that entanglement of such concepts with a subject and the need for using multiple word embeddings in describing more complex concepts hinders current inversion methods. We proposed Lego that effectively inverts such personalized concepts from as few as 4 example images by disentangling concepts from subjects with a *Subject Separation* step and further enables defining concepts with multiple embeddings by using a *Context Loss* that guides each embedding towards a meaningful place in the textual embedding space.

Acknowledgements. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023)
2. Amazon: Amazon mechanical turk. https://www.mturk.com/ (2023), https://www.mturk.com/
3. Avrahami, O., Hertz, A., Vinker, Y., Arar, M., Fruchter, S., Fried, O., Cohen-Or, D., Lischinski, D.: The chosen one: Consistent characters in text-to-image diffusion models. arXiv preprint arXiv:2311.10093 (2023)
4. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers (2022)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
8. Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., Allahbakhsh, M.: Quality control in crowdsourcing. ACM Computing Surveys (CSUR) **51**, 1 – 40 (2018), https://api.semanticscholar.org/CorpusID:2700101
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
11. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021)
12. Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. Advances in Neural Information Processing Systems **36** (2024)
13. Face, H.: Diffusers library. https://huggingface.co/docs/diffusers/index (2023), https://huggingface.co/docs/diffusers/index
14. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032 (2022)
15. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
16. 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 preprint arXiv:2208.01618 (2022)
17. 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 Transactions on Graphics (TOG) **42**(4), 1–13 (2023)

18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
19. 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. pp. 10696–10706 (2022)
20. Hendricks, L.A., Nematzadeh, A.: Probing image-language transformers for verb understanding. arXiv preprint arXiv:2106.09141 (2021)
21. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
24. Huang, Z., Wu, T., Jiang, Y., Chan, K.C., Liu, Z.: Reversion: Diffusion-based relation inversion from images. arXiv preprint arXiv:2303.13495 (2023)
25. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
26. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (2022)
27. 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. pp. 1931–1941 (2023)
28. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access **8**, 193907–193934 (2020). `https://doi.org/10.1109/access.2020.3031549`, `http://dx.doi.org/10.1109/ACCESS.2020.3031549`
29. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
30. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
31. 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)
32. 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)
33. Momeni, L., Caron, M., Nagrani, A., Zisserman, A., Schmid, C.: Verbs in action: Improving verb understanding in video-language models. arXiv preprint arXiv:2304.06708 (2023)
34. 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)

35. 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)
36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
37. Park, J.S., Shen, S., Farhadi, A., Darrell, T., Choi, Y., Rohrbach, A.: Exposing the limits of video-text models through contrast sets. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3574–3586 (2022)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
39. 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)
40. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
41. Rao, H., Huang, S.W., Fu, W.T.: What will others choose? how a majority vote reward scheme can improve human computation in a spatial location identification task. In: AAAI Conference on Human Computation & Crowdsourcing (2013), `https://api.semanticscholar.org/CorpusID:8001594`
42. 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)
43. 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)
44. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
45. 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)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
47. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
49. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
50. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. ACM Transactions on Graphics **42**(6), 1–13

(Dec 2023). https://doi.org/10.1145/3618315, http://dx.doi.org/10.1145/3618315
51. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
52. Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., Chang, S.: Uncovering the disentanglement capability in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1900–1910 (June 2023)
53. 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 preprint arXiv:2206.10789 **2**(3), 5 (2022)
54. Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., Kale, A., Faieta, B.: Multimodal contrastive training for visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6995–7004 (June 2021)
55. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022)
56. 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)