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

- We go over Lego's limitation and ethical use of personalized generative models in Section 7
- Please refer to Section 8 to see the effect of increasing the number of reference images for performing Lego.
- In Section 9 we show that the learned subject embedding by Lego can also be used and it performs similar to TI.
- To see **Lego inverted concepts applied to more subjects**, please see Figure 11.
- If you are interested in **details of the experimental design** such as text templates and how $\mathcal{P}$ and $\mathcal{N}$ were chosen for each concept, please refer to Section 11.
- For details on **the prompts used in generating concepts using the baseline and details on the VQA study**, please see Section 12.
- 200 images generated using Lego and 200 images generated using the baseline are shown in Figures 16 to 29.
- We show how ELITE [51] and ControlNet [56] perform in inverting general concepts in Section 13.
- The full results for our ablation study can be found in Section 15.

7 Limitations and Ethical Statement

Lego struggles to invert concepts beyond the capabilities of the backbone, like facial expressions with earlier versions of Stable Diffusion. Our research is dedicated to ethical and responsible data use, emphasizing the need for socially responsible applications in personalized visual media.

8 The Effect of Dataset Size

In our experiments, we only used 4 images per experiment; 2 reference images of the subject only and 2 reference images of the concept applied to the subject. We trained Lego for the "closed eyes" and "raised arms" concept with 2, 4 & 8 concept images and 2 subject-only images. We then generated 100 images for each concept. We then used our LLM-VQA evaluation setting to evaluate Lego's performance in each of these settings. For the concept of "closed eyes", Lego achieved an accuracy of 75% when 2 concept images were shown to Lego. This increased to 83% for both experiments with 4 and 8 concept reference images. Similarly, for the concept of "raised arms", an accuracy of 78% was achieved with only 2 concept images. Using 4 and 8 concept images increased the accuracy to 80% and 81% respectively.

9 Lego's Ability in Subject Synthesis

While Lego's main objective is to learn a good representation for synthesizing a concepy, it does so with a Subject separation step that also learns an embedding to represent the sibject. We conducted experiments to see whether the learned subject embedding performs as well as TI in synthesizing the subject. We generated 40 images of 2 subjects with TI and Lego. In a user preference study with 5 participants, TI was preferred 55% of the time over Lego (45%), showing similar capabilities in generating the subject with TI. Two such examples are shown in Figure 10.

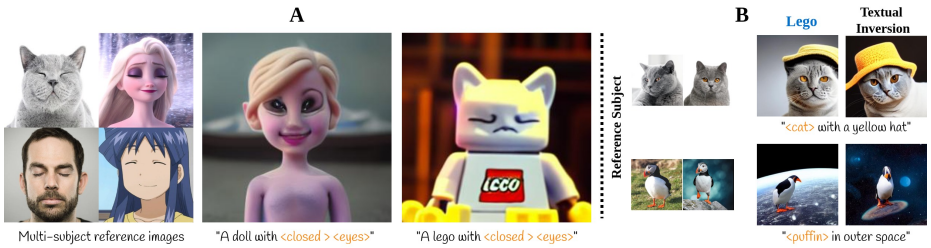

Fig. 10: (**A**) Appearance leakage in TI for inversion of "closed eyes" while using multiple subject reference images. (**B**) Lego and TI show similar capabilities in synthesizing objects cat and bird.

10 Additional Results

In Figure 11, we show the results of Lego learned concepts, applied to four additional subjects.

11 Experiment Design

In this section, we go over the design of each concept inversion experiment, including the template of prompts used and selecting the appropriate positive ($\mathcal{P}_i$) and negative ($\mathcal{N}_i$) set of words for each concept embedding.

11.1 Numerical Concepts

In our work, we inverted concept numbers 3, 4 and 5 by using example images of 1 Lego figurine as $\mathcal{I}_{\overline{C}}$ and images of 3, 4, and 5 Lego figurines as $\mathcal{I}_C$, respectively. Figure 12 shows sample text templates used for inverting such concepts. We dedicated an embedding $<subj>$ for learning the subject appearance and an

embedding $<cpt>$ to learn the concept. For all numerical concepts, a single concept embedding was used.

To form the positive and negative sets of words that were used to calculate our context loss, we constructed the positive set $\mathcal{P}$ to be the digit and word description of the desired number and further constructed the negative set $\mathcal{N}$ to be the digit and word descriptions of a few neighbouring numbers. For instance, to learn the concept of 4 from the Lego figurines in Figure 12, we set $\mathcal{P} = \{\text{four}, 4\}$ and $\mathcal{N} = \{2, 3, 5, 6, \text{two}, \text{three}, \text{five}, \text{six}\}$.

11.2 Remaining 7 Concepts

In Figure 13, we show template samples for the remaining 7 concepts that were explored in this paper. We skip the subject-only templates as they follow the same structure as the one we showed in Figure 12. To the right of each concept's template, we provide each embedding's corresponding $\mathcal{P}$ and $\mathcal{N}$ that we used in our experiments.

In order to select $\mathcal{P}$ and $\mathcal{N}$, we used a few synonyms for the effect we are trying to invert. In the case of numerical concepts, our negative set was the neighbouring numbers due to the fact that they are in close embedding distance to the concept number. This is however not the case in many examples. For instance, for the concept " closed eyes", the positive set for one of the embeddings has the word "closed". The antonym word "open" however is not close to the embedding of "closed". Our experiments showed that in such cases, the negative set's words is not as important as the positive set, yet one should still select words based on the antonym scheme as selecting arbitrary words can cause the embedding to diverge from reaching a suitable place in the text embedding space.

We suggest following our sanity check step when using Lego to invert arbitrary concepts; during optimization, for each concept embedding, we monitor the 10 most and least similar words to the embedding. This allows you to see if your embedding is indeed getting close to a set of words that describe the concept. They could also suggest words for building the negative set if the embedding is also getting close to undesired words that you had not thought of before.

12 LLM and Human Preference Study

In this section, we provide the 200 images generated with Lego and LDM for each concept in our study. Both the VQA LLM (Flan) and Human preference (Mechanical Turk) study preferred Lego generated concepts over the baseline (LDM with natural language input).

For our evaluation with the VQA model, we experimented with asking different types of questions based on the ground truth exemplar images. "Yes" or "No" questions about the concept being present in the image or not led to the highest correct answers based on the example images. Hence, we asked "Yes" or "No" questions about each image and according to the LLM, Lego generated concepts

had more images aligned with each concept compared to the baseline, across all concepts (see Table 2).

For each concept in the Human preference study, 10 sets of questions were created, each containing 20 pairs of images, one generated using Lego's inversion of the concept synthesized through an LDM backbone and one generated using natural language description (see next paragraph for details) of the concept, synthesized through LDM. For each pair of images, users were prompted to pick one image between the two that best represents the concept (users were given both the exemplar images and text description of the concept for each question - see Figure 14 for a snapshot of the user study interface for the concept of *walking on rope*). We selected the image with majority user vote [8, 41] for each pair as the preferred concept representation. Lego was preferred over the baseline in all concepts as shown in Table 2.

In Figures 16 to 29 show the 200 images generated for each concept using Lego and natural language. With prompt tuning being an important aspect of getting the desired outcome from generative models, for each concept in the baseline experiments, we described the concept in 4 different natural language prompts, with LDM generating 50 images per prompt. Below we provide the prompts used for each concept.

We applied the concept of ***frozen in ice*** to "toy bear" (see Figure 16) and we used the following prompts for LDM; 1) "Photo of a toy bear frozen in ice", 2) " a toy bear stuck in a block of ice", 3) "a toy bear that is frozen in a block of ice" and 4) "photo of a toy bear that is stuck in ice" (see Figure 17).

We applied the concept of ***burnt and melted*** to "toy bear" (see Figure 18) and we used the following prompts for LDM; 1) "Photo of a toy bear that is burnt and melted", 2) "photo of a melted and burnt toy bear", 3) "a toy bear that is burnt and charred" and 4) "photo of a burnt and destroyed toy bear" (see Figure 19).

We applied the concept of ***closed eyes*** to "lion" (see Figure 20) and we used the following prompts for LDM; 1) "a lion with closed eyes", 2) "photo of a lion with its eyes closed", 3) "photo of a lion with its eyes shut" and 4) "photo of a lion that has closed eyes" (see Figure 21).

We applied the concept of ***smiley emoji face*** to "fluffy monster" (see Figure 22) and we used the following prompts for LDM; 1) "Photo of a fluffy monster with a smiley emoji face", 2) "photo of a fluffy monster with a yellow smiley emoji head", 3) "a fluffy monster with a yellow smiley emoji face" and 4) "photo of a fluffy monster having a yellow smiley emoji mask" (see Figure 23).

We applied the concept of ***raised arms*** to "toy bear" (see Figure 24) and we used the following prompts for LDM; 1) "Photo of a toy bear with its arms raised", 2) " photo of a toy bear with arms up", 3) "photo of a toy bear with its hands up in the air" and 4) " photo of a toy bear having raised its arms" (see Figure 25).

We applied the concept of ***crumpled and crushed*** to "toy car" (see Figure 26) and we used the following prompts for LDM; 1) "Photo of a toy car that is crushed and crumpled", 2) " photo of a crushed and destroyed toy car", 3) "photo

of a crumpled toy car" and 4) " photo of a crushed and squeezed toy car" (see Figure 27).
We applied the concept of ***walk on rope*** to "zebra" (see Figure 28) and we used the following prompts for LDM; 1) "Photo of a zebra walking on a rope", 2) " photo of a zebra balancing on a tightrope", 3) "a zebra walking on a tightrope" and 4) " photo of a zebra walking on a wire" (see Figure 29).

13 ControlNet and ELITE

recall that ELITE [51] trained mapping networks on a large vision dataset in order to skip the embedding optimization step of Textual inversion by directly learning to map an image's subject to its corresponding embedding. Similar to TI, ELITE is an appearance based inversion method and also requires an input mask of the subject to be inverted. In Figure 15 we show ELITE's attempt at inverting concepts of *walking on a rope, smiley emoji face, frozen in ice* and *crumpled and crushed*.
ControlNet [56] is a powerful method for adding conditional control to T2I Diffusion models; such as pose and layout. In Figure 15 we condition Stable Diffusion on the edge map of example images of our concepts and prompt the model to follow these outlines. This is too restrictive for general concepts and as seen in the four concept examples, the results are not satisfactory.

14 All Lego and Baseline Images Used In User + LLM Study

In Figures 16 to 29, we show all images for the concepts we used in our User + LLM study.

15 Complete Ablation Results

In Section 5.5 we showed a few ablation results for the concept of *burnt and melted*. In Figures 30 and 31, we show the full ablation results for the single and multi subject *burnt and melted* concept. In the single subject experiment, we used $\mathcal{P}$ words that were more focused on **burnt** and **destroyed** whereas in the multi-subject experiment we used words focused on **melted** and **liquefy**. The effect is clear in Lego's inversion results based on the choice of positive word selection. We encourage the user to experiment with the words that they would like to prominently see in the result concept. Figures 32 and 33 show the ablation results for learning the concepts of *closed eyes* and *walking on rope* and applying them to "batman" and "a zebra" respectively.

The ablation for the three concepts clearly show the effect of **Subject Separation** and **Context Loss** on the concept's representation. In the single subject melting examples, we can see Rubik's cube's features in the toy bear and in the

multi-subject scenario, both the features of the Rubik's cubes and Toy Story's Woody can be seen in the toy bears (notice Woody's eyes in some of the bears). Furthermore, in the *closed eyes* example, the cat's features can be found in what is mean to represent Batman's face. In the *walking on rope* example, the figure from the example images is present instead of "a woman". The effect of context loss can be seen by comparing Lego generated concepts with concepts generated by performing subject separation and not context loss. Not performing context loss leaves most images unfaithful in representing the concept.

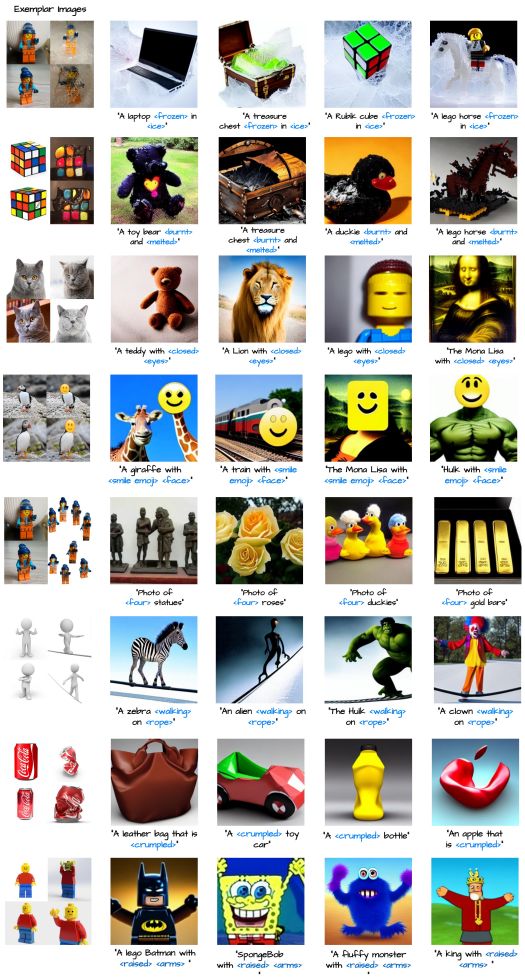

Fig. 11: We show Lego learned concepts applied to more subjects. please zoom in to see details of images.

"A photo of a <subj>"
"A nice depiction of <subj>"
"A good photo of the <subj>"
. . .

"A photo of <cpt> <subj>"
"A nice depiction of <cpt> <subj>"
"A good photo of <cpt> <subj>"
. . .

Fig. 12: Sample templates for inverting numerical concepts. Single subject images will be make up $\mathcal{I}_{\overline{C}}$ and the templates will only have the $<subj>$ (left) while images with as many subjects as we are trying to invert the number for will make up $\mathcal{I}_C$ with the template having both $<subj>$ and $<cpt>$ embeddings (right).

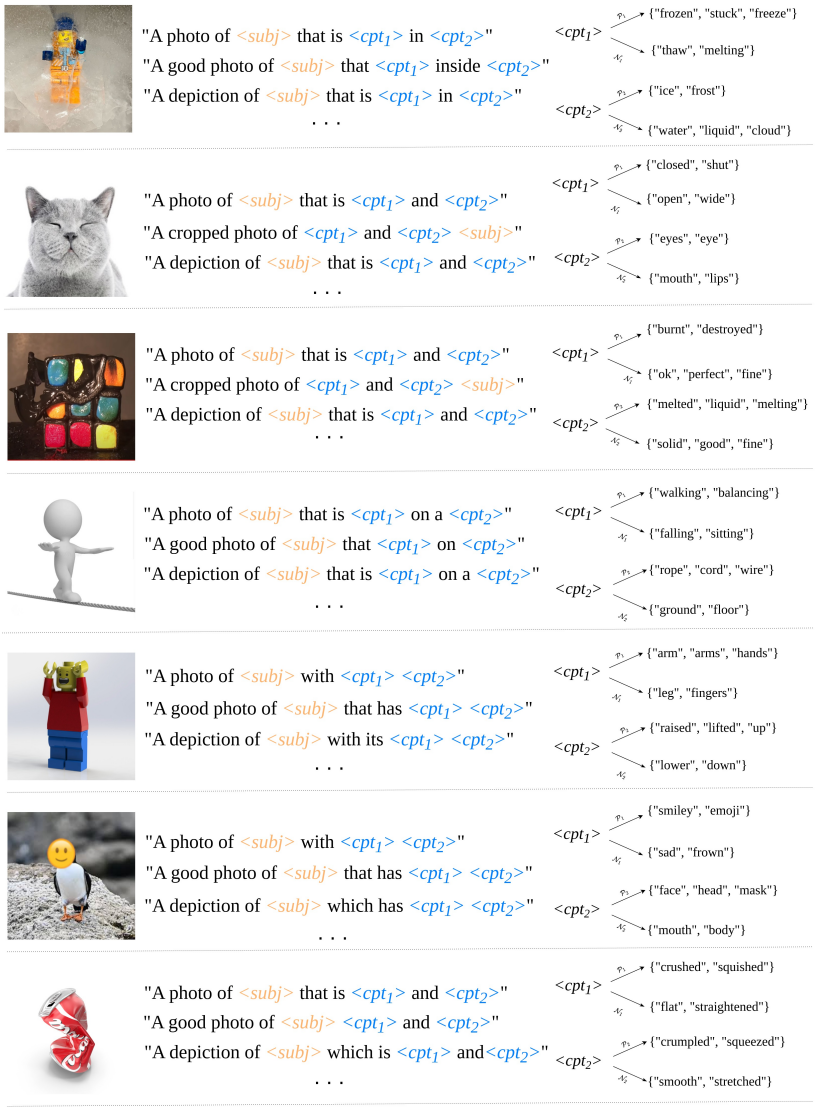

Fig. 13: Template samples for the non-numerical concepts explored in this work. All 7 concepts use two word embedding and we show the positive and negative words for each embedding.

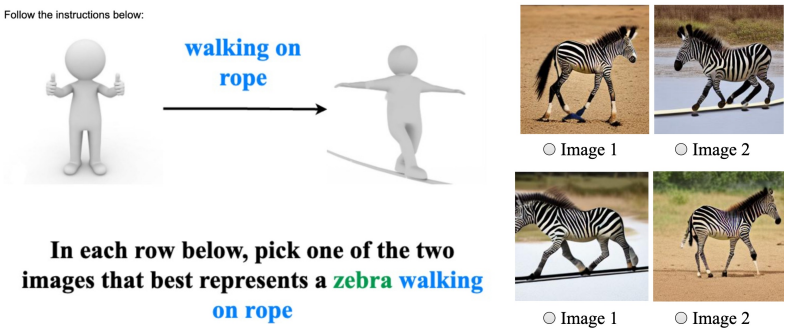

Fig. 14: The users were shown the image with instructions (left) on the concept they should be looking for in each pair of images. And below the instruction, 20 questions, each including 2 images (right). were given to users where they had to pick one of the two. The order of Lego and baseline generated concepts was randomized.

Fig. 15: Results of ControlNet and ELITE for inverting concepts of *walking on a rope, smiley emoji face, frozen in ice* and *crushed and crumpled*

Fig. 16: Lego generated images with LDM backbone for the concept of *frozen in ice*, learned from the Lego figurine frozen in ice example images, applied to a toy bear.

Fig. 17: LDM generated images using 4 different prompts, describing a toy bear that is frozen in ice.

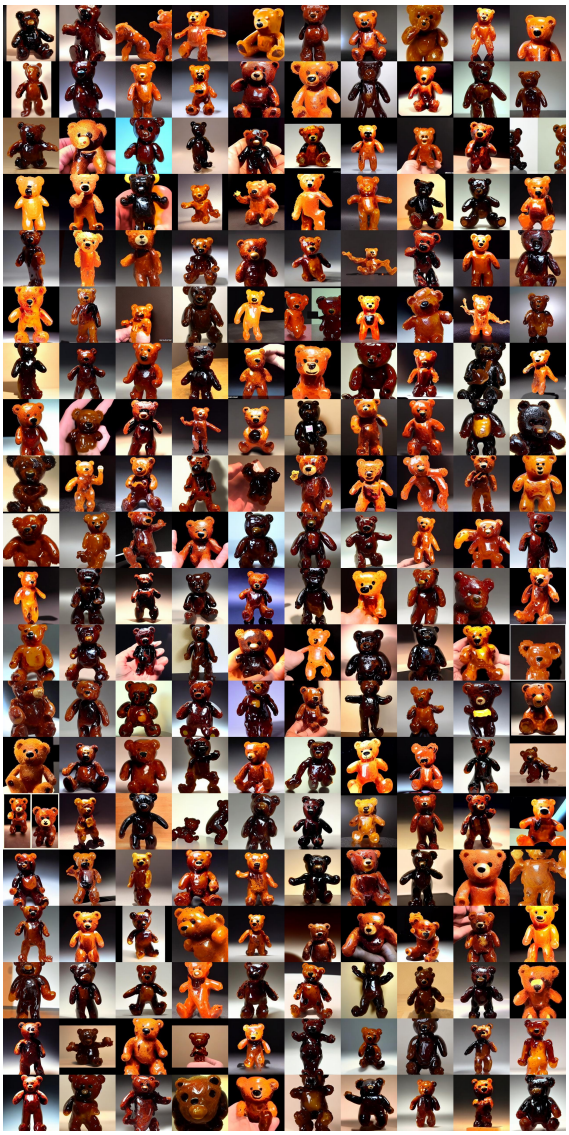

Fig. 18: Lego generated images with LDM backbone for the concept of *burnt and melted*, learned from the burnt and melted Rubik's cube example images, applied to a toy bear.

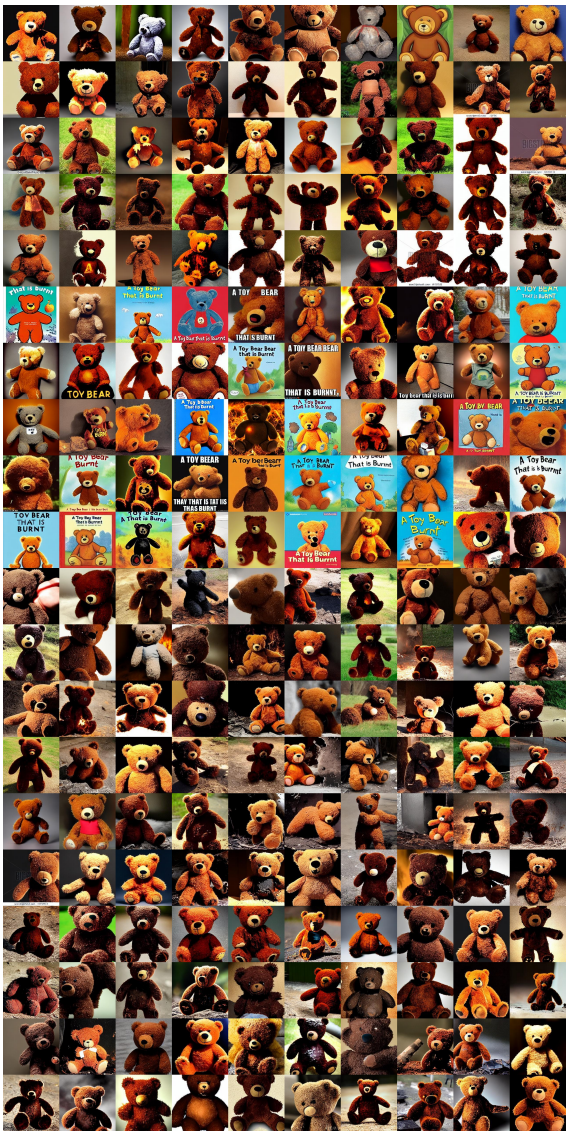

Fig. 19: LDM generated images using 4 different prompts, describing a toy bear that has been burnt and melted.

Fig. 20: Lego generated images with LDM backbone for the concept of *closed eyes*, learned from the cat with closed eyes example images, applied to a lion.

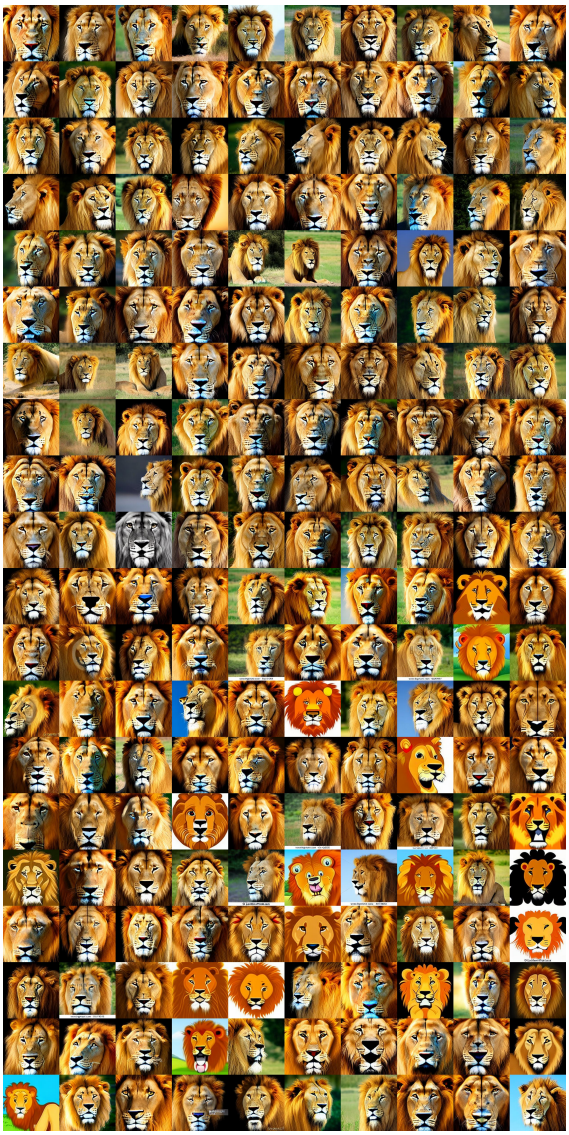

Fig. 21: LDM generated images using 4 different prompts, describing a lion with its eyes closed.

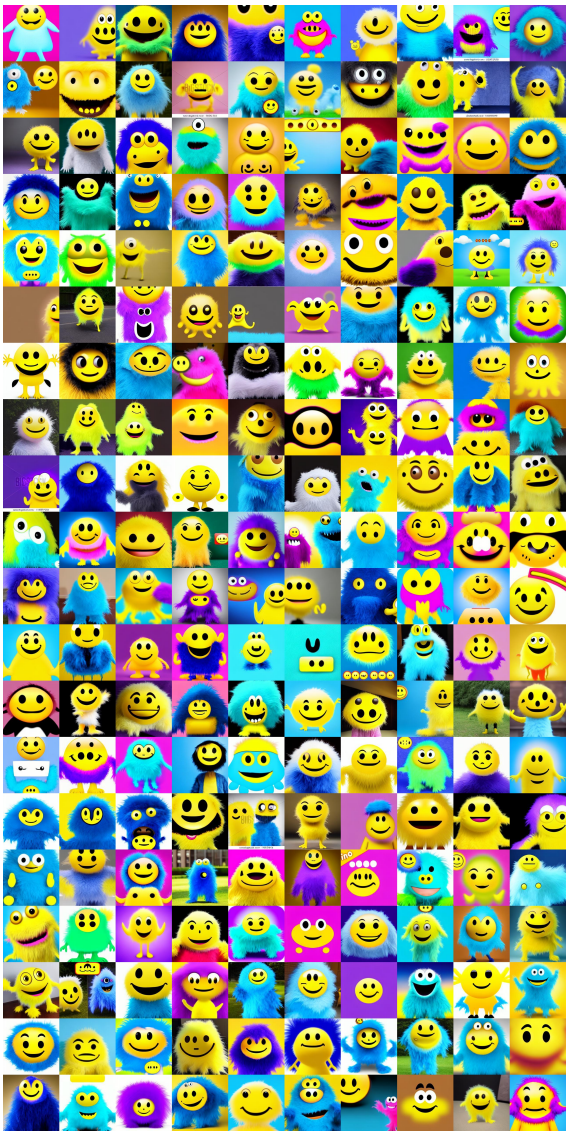

Fig. 22: Lego generated images with LDM backbone for the concept of *smiley emoji face*, learned from the bird with smiley emoji face example images, applied to a fluffy monster.

Fig. 23: LDM generated images using 4 different prompts, describing a fluffy monster with a yellow smiley emoji face.

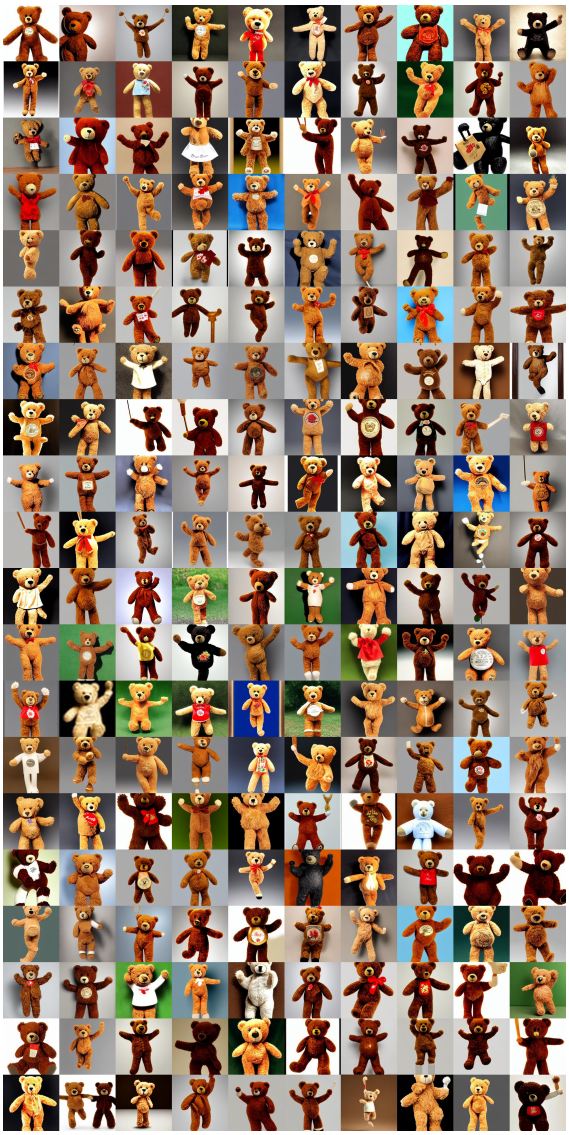

Fig. 24: Lego generated images with LDM backbone for the concept of *arms raised*, learned from the Lego figurine with its arms raised example images, applied to a toy bear.

Fig. 25: LDM generated images using 4 different prompts, describing a toy bear with its arms raised.

Fig. 26: Lego generated images with LDM backbone for the concept of *crushed and crumpled*, learned from the crushed and crumpled soda can example images, applied to a toy car.

Fig. 27: LDM generated images using 4 different prompts, describing a toy car that is squished and crumpled.

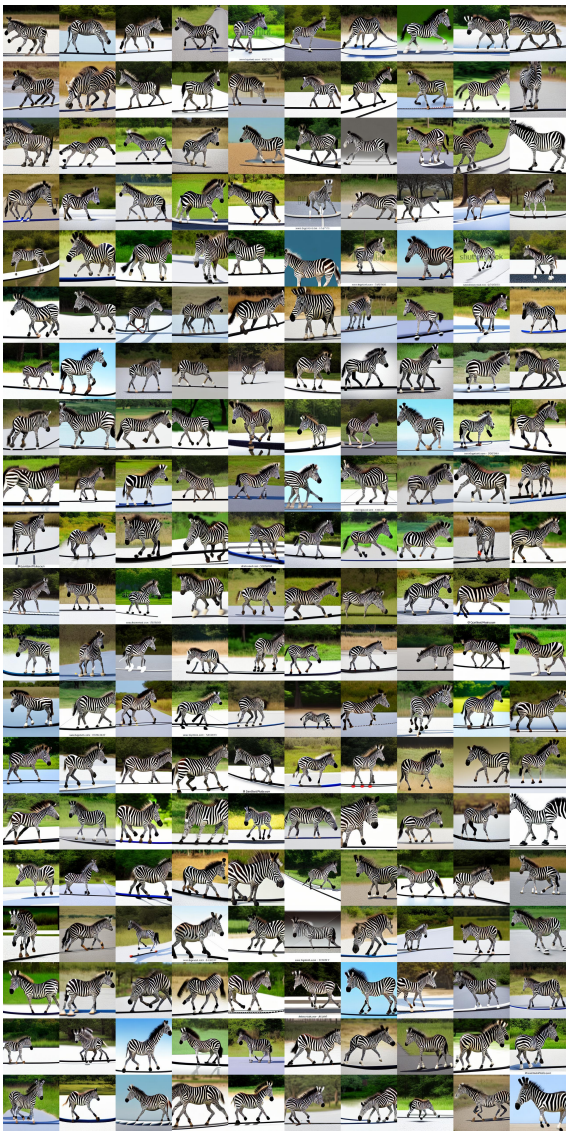

Fig. 28: Lego generated images with LDM backbone for the concept of *walking on rope*, learned from the figure walking on a rope example images, applied to a zebra.

Fig. 29: LDM generated images using 4 different prompts, describing a zebra that is walking on a rope.

Fig. 30: Ablation results for *burnt and melted* concept from single subject example images applied to the subject "toy bear". Please zoom in to see details.

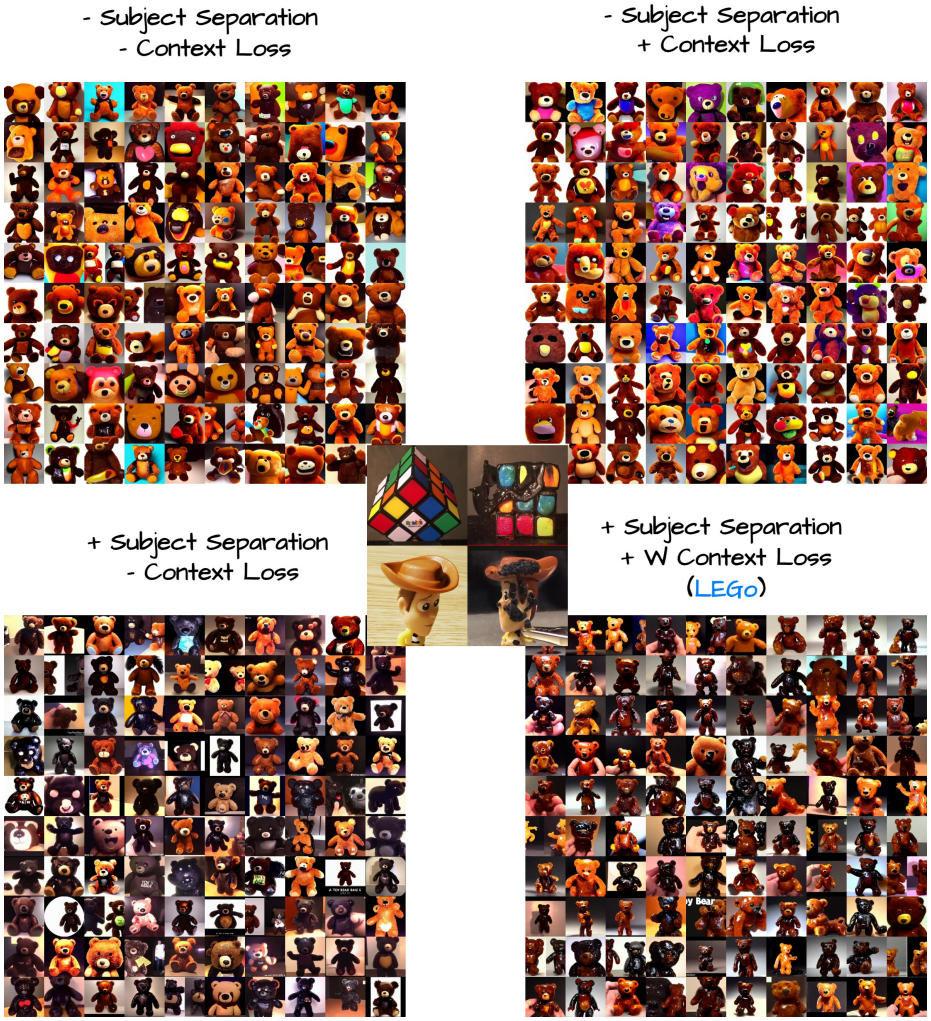

Fig. 31: Ablation results for *burnt and melted* concept from multi subject example images applied to the subject "toy bear". Please zoom in to see details.

Fig. 32: Ablation results for *closed eyes* concept applied to the subject "batman". Please zoom in to see details.

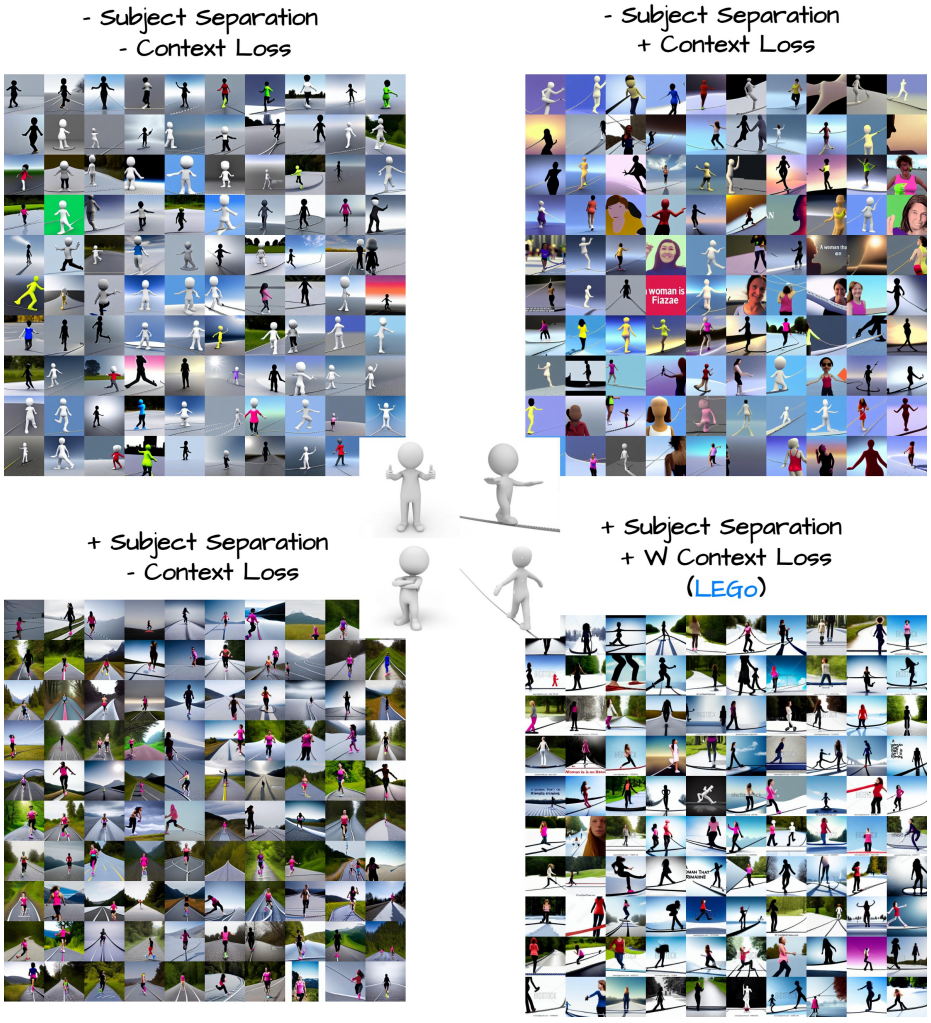

Fig. 33: Ablation results for *walking on rope* concept applied to the subject "a woman". Please zoom in to see details.