



中山大學
SUN YAT-SEN UNIVERSITY



AttnDreamBooth: Towards Text-Aligned Personalized Text-to-Image Generation

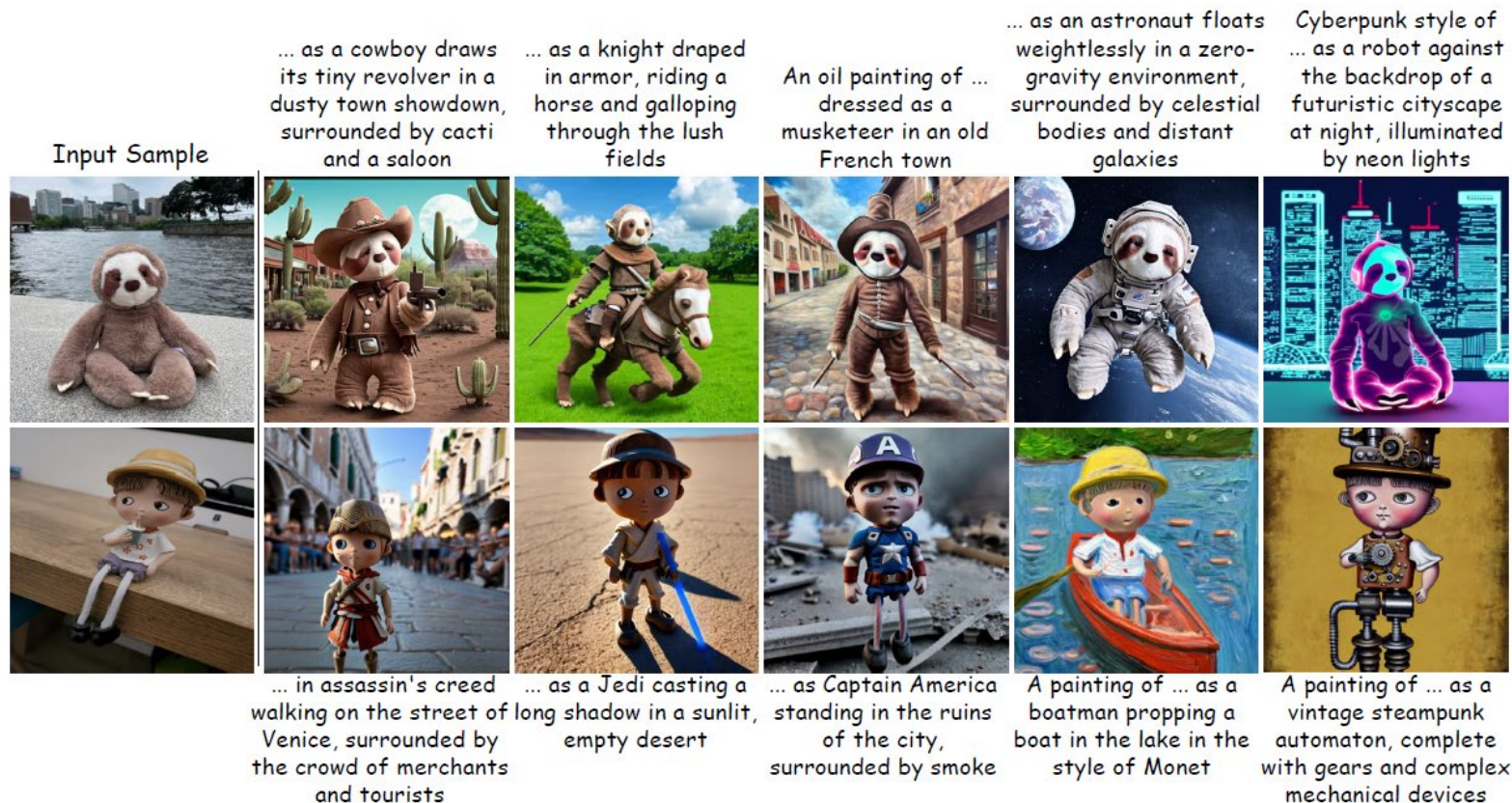
Lianyu Pang¹, Yin Jian¹, Baoquan Zhao¹, Feize Wu¹, Fu Lee Wang², Qing Li³, Xudong Mao^{1,*}

¹Sun Yat-sen University ²Hong Kong Metropolitan University

³The Hong Kong Polytechnic University

- Introduction
- Method
- Experiment
- Conclusion

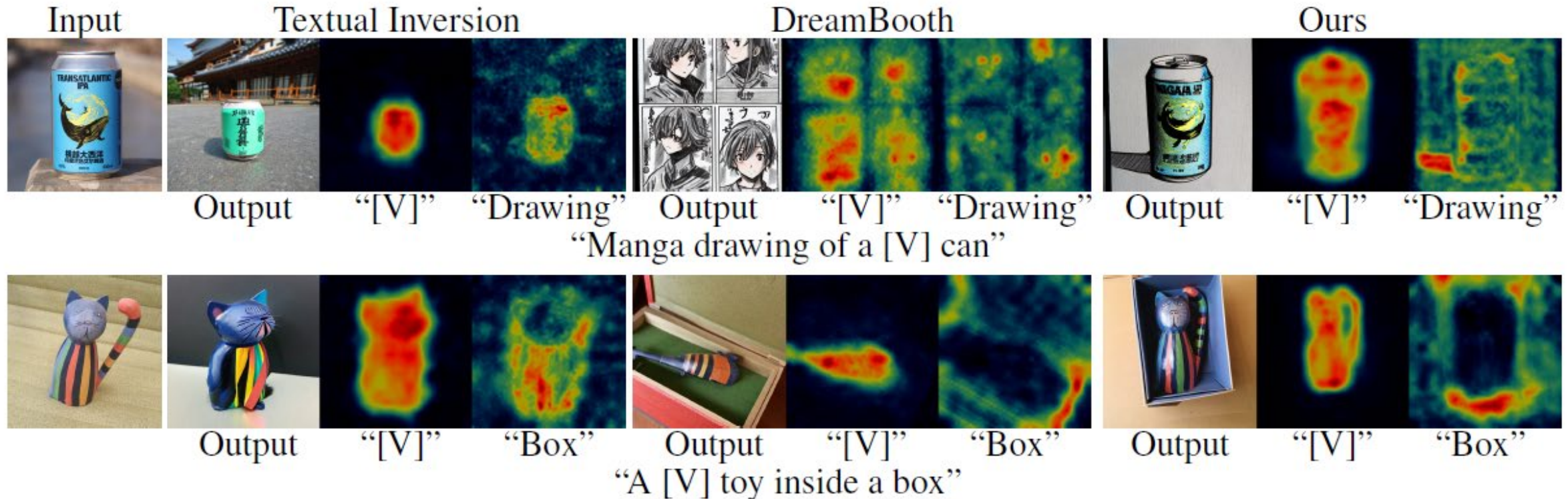
- Text-to-image personalization is the task of customizing a pre-trained diffusion model to produce images of user-provided concepts in novel scenes or styles.
- Current personalization techniques struggle to balance the trade-off between **identity preservation** and **text alignment**.
- Our method achieves superior performance in terms of identity preservation and text alignment compared to the baselines.



Introduction

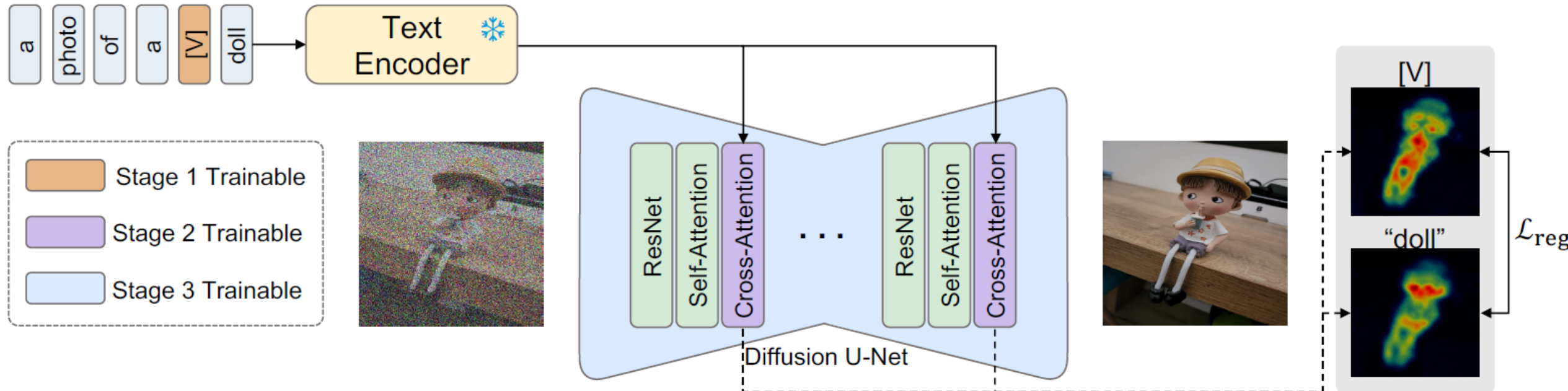
4

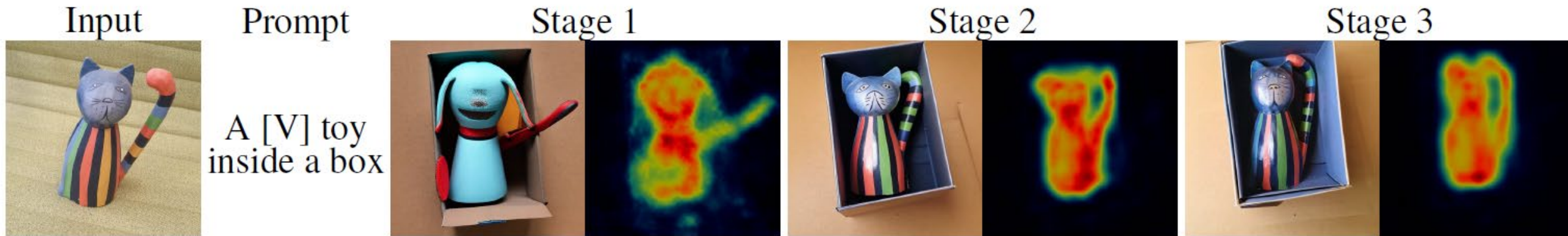
- The two principal methods **Textual Inversion** and **DreamBooth** encounter distinct challenges when integrating the learned concept into novel prompts.
- Textual Inversion tends to **overfit** the textual embedding of the learned concept, resulting in incorrect attention map allocations to other tokens (e.g., “drawing” or “box”). In contrast, DreamBooth appears to **overlook** the learned concept, producing images primarily based on other tokens.
- These issues can be attributed to the incorrect learning of embedding alignment for the new concept, i.e., the embedding of the new concept is not functionally compatible with the embeddings of existing tokens.



- Based on these observations, our approach aims to properly learn not only the subject identity but also the embedding alignment and the attention map for the new concept. Our key insights are as follows:
 1. First, while **Textual Inversion** often fails to capture the subject identity and tends to overfit the embedding alignment for the new concept, it can effectively learn the embedding alignment and a coarse cross-attention map in the very early stages of optimization.
 2. Second, although **DreamBooth** fails to learn the embedding alignment, it can accurately capture the subject identity.
- We set the training prompt as “a photo of a [V] [super-category]”, and introduce a cross-attention map regularization term, which serves two purposes.
 1. First, since the new concept and its super-category belong to the same object category, the attention map of the super-category token can serve as a reference for the new concept.
 2. Second, since [V] and [super-category] are used together to describe the new concept when integrating it into new prompts, the attention maps of [V] and [super-category] should refer to the same region

- We propose to decompose the personalization process into three training stages:
 1. Learning the embedding alignment
 2. Refining the attention map
 3. Acquiring the subject identity
- Furthermore, we introduce a cross-attention map regularization term to enhance the learning of the attention map





- The generations along with the attention maps of “[V]” for each stage.
 1. In stage 1, the model properly aligns the embedding of [V] with other tokens, “inside a box”, but learns a very coarse attention map and subject identity.
 2. In stage 2, the model refines the attention map and subject identity.
 3. In stage 3, the model accurately captures the identity of the concept.

Experiment

8

● Dataset

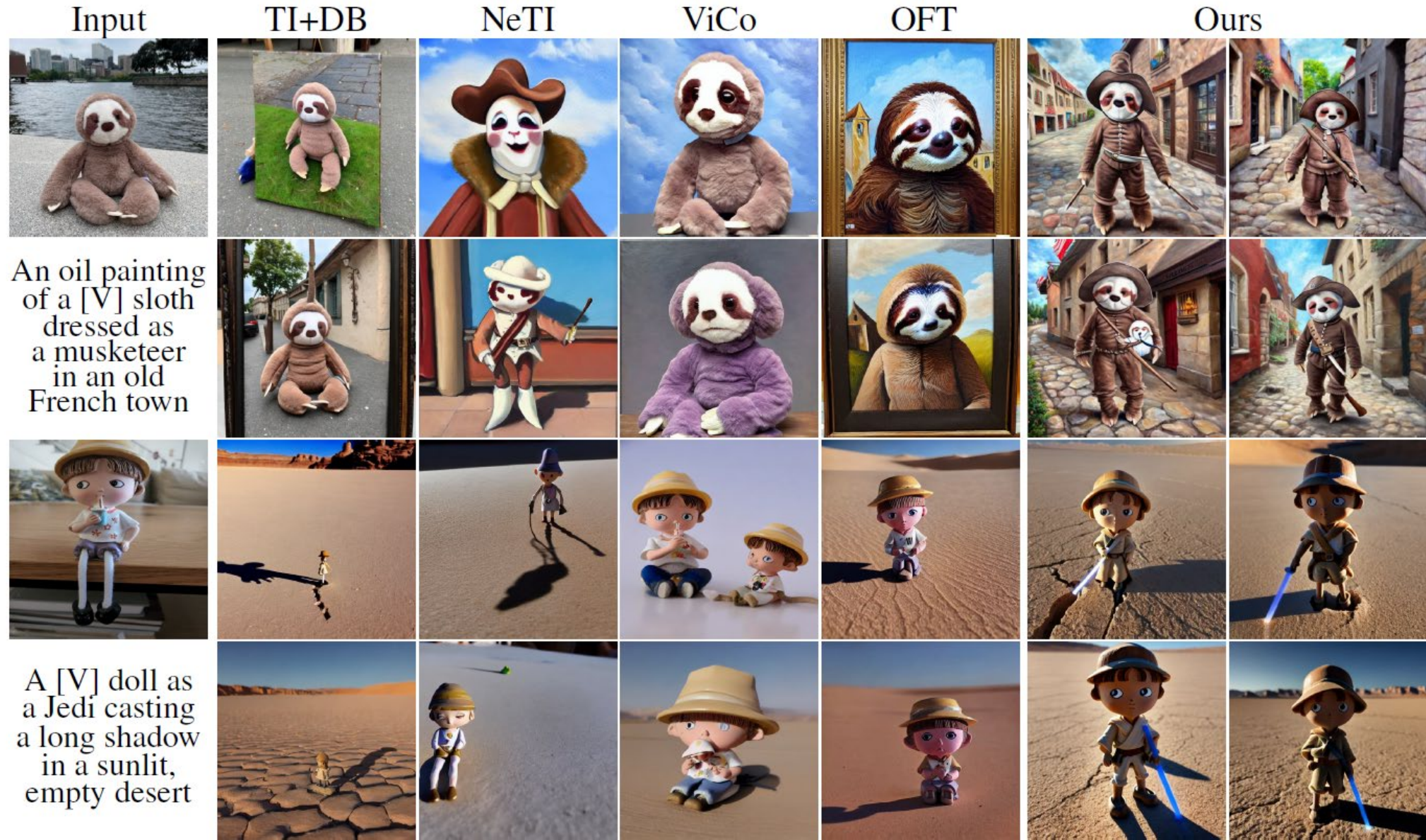
- We collect 22 concepts from **Textual Inversion** and **DreamBooth**.
- We use a set of 24 text prompts for the quantitative evaluation.

● Metric

- **Identity preservation:** measured by the cosine similarity between the CLIP embeddings of generated and real images
- **Text alignment:** measured by the cosine similarity between the CLIP embeddings of generated images and their corresponding prompts.
- Each method is evaluated using 24 text prompts, generating 32 images per prompt.

Table 1: **Quantitative comparisons.** “Identity” denotes the identity preservation, and “Text” denotes the text alignment.

Methods	Identity $\uparrow$	Text $\uparrow$
TI+DB [24, 71]	0.7017	0.2578
NeTI [1]	0.6901	0.2522
ViCo [30]	0.7507	0.2106
OFT [64]	0.7257	0.2445
Ours-fast	<u>0.7268</u>	<u>0.2536</u>
Ours	0.7257	0.2532



Input



TI+DB



NeTI



ViCo



OFT



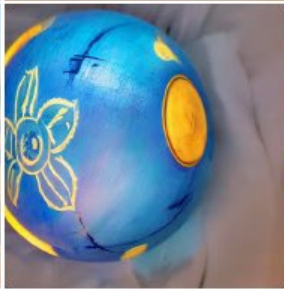
Ours



A [V] fluffy
walking in
the rainy
streets



Light seeps out
from the inside
of a clear [V]
teapot under the
moonlight on
a serene beach





Input Sample



A black [V]
toy wearing
sunglasses on
the beach



A [V] toy wearing
a chef hat in a
kitchen with meat
and vegetables on
the table



A [V] toy
wearing a
police cap in
a police car



A [V] toy as a
priest in blue robes
in the cathedral



Input Sample



A purple [V]
furby under the
mystical aurora
borealis in a
remote Arctic
landscape



A red [V] furby
against the back-
drop of a futuri-
stic cityscape at
night illuminated
by neon lights



A [V] furby amidst
a bustling street
market surrounded
by vibrant colors
and textures



A black [V]
furby bathed in
the golden light
of sunset at a
serene beach



Input Sample



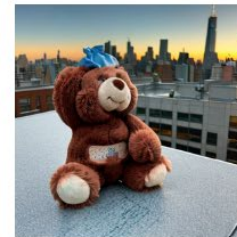
A [V] bear atop
a high cliff
overlooking
stormy seas



A [V] bear
surrounded by
fluttering butterflies
in a meadow



A [V] bear in
the reflection
of a cracked
antique mirror

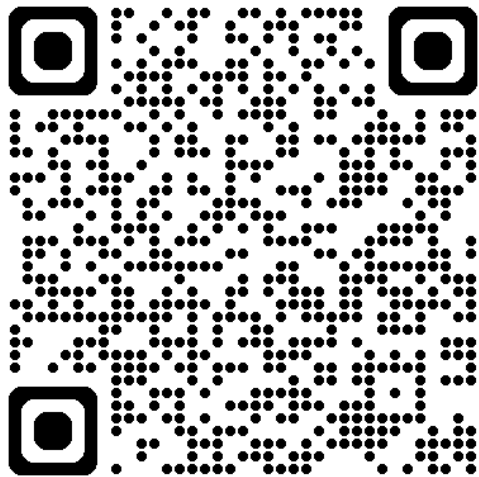


A [V] bear
perched on a
city rooftop
at sunset

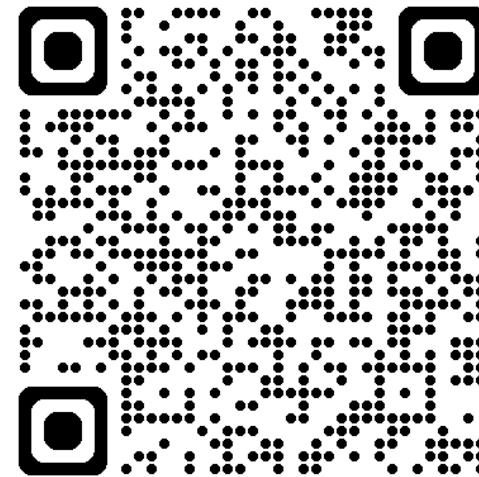


A [V] bear
floating in the
weightlessness
of space

- We identified and analyzed the embedding misalignment issue encountered by **Textual Inversion** and **DreamBooth**.
- Our proposed method, named **AttnDreamBooth**, addresses this issue by decomposing the personalization process into three stages: learning the embedding alignment, refining the attention map, and acquiring the subject identity.
- **Our method** enables identity-preserved and text-aligned text-to-image personalization, even with **complex prompts**.



Arxiv



Github



中山大學
SUN YAT-SEN UNIVERSITY



Thanks for your attention