An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Sangyun Lee

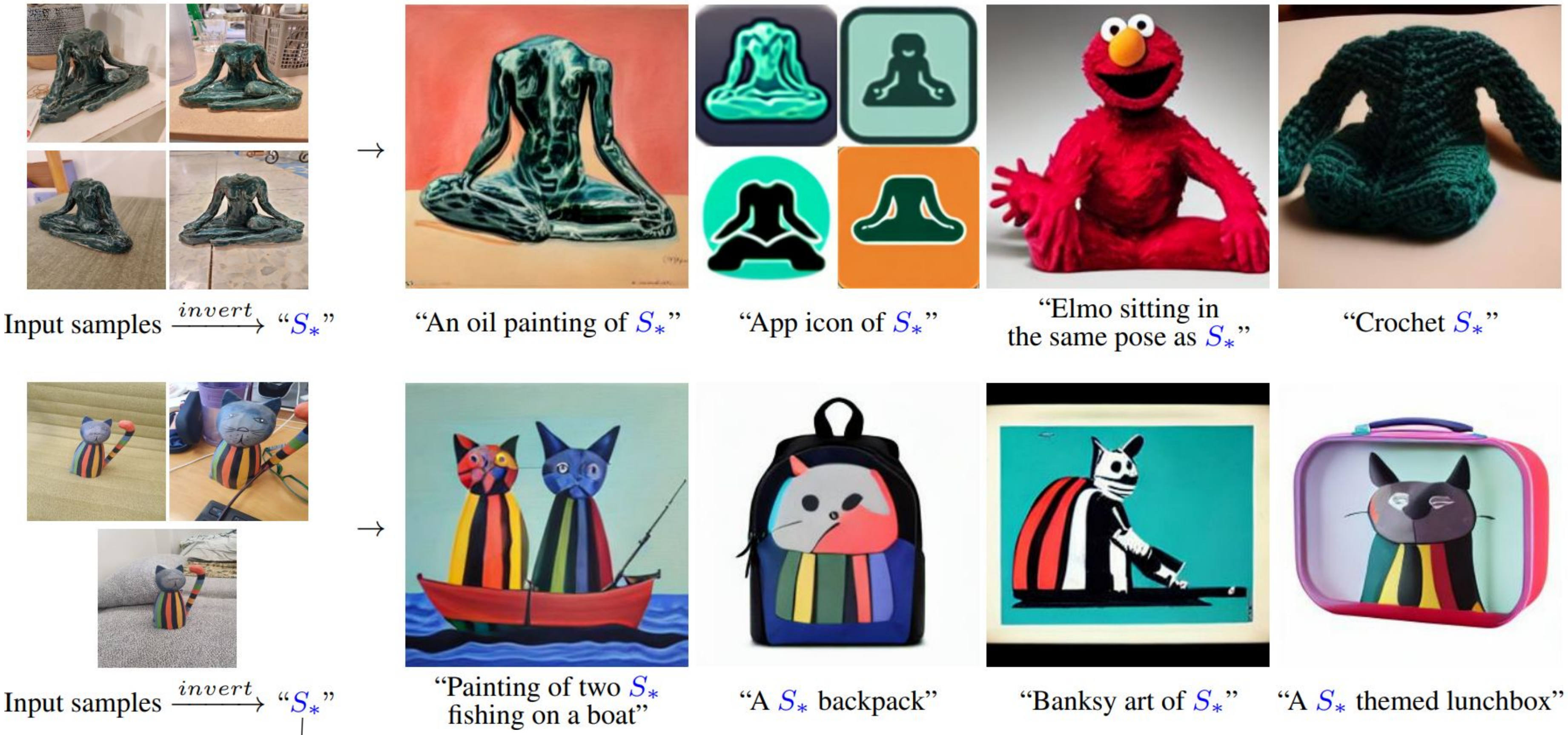
a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

panda mad scientist mixing sparkling chemicals, artstation

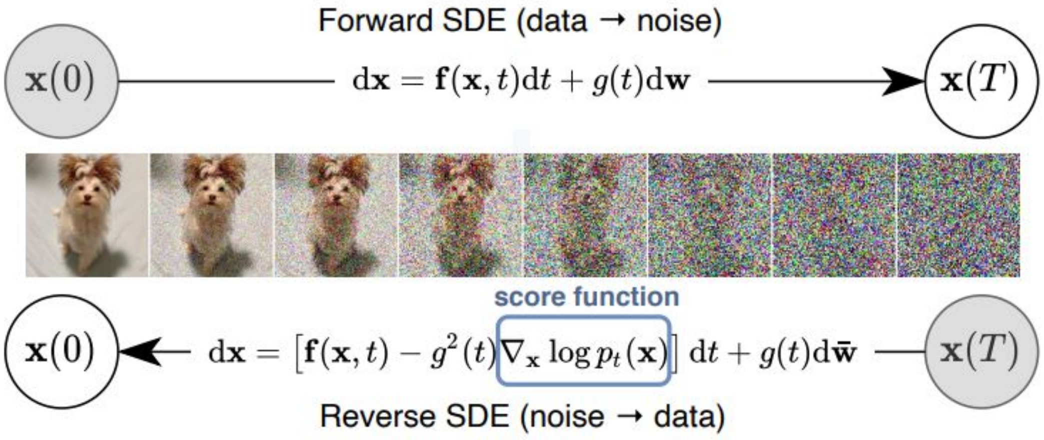
a corgi's head depicted as an explosion of a nebula

It is difficult to fully control the generative process using text prompt only.



cat doll wearing a vertical striped clothing, where stripes are black, yellow, red, and blue, and …… find this via optimization!

Diffusion models



Song et al., 2020

Ho et al., 2020

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w}.$$

Denoising AEs are score estimators

$$\arg\min_{\theta} \mathbb{E}_i\{\lambda(i)\mathbb{E}_{q_0(\mathbf{x})}\mathbb{E}_{q(\mathbf{x}_i|\mathbf{x}_0)}[||\mathbf{s}_\theta(\mathbf{x}_i, i) - \nabla_{\mathbf{x}} \log q_i(\mathbf{x}_i|\mathbf{x}_0)||_2^2]\},$$

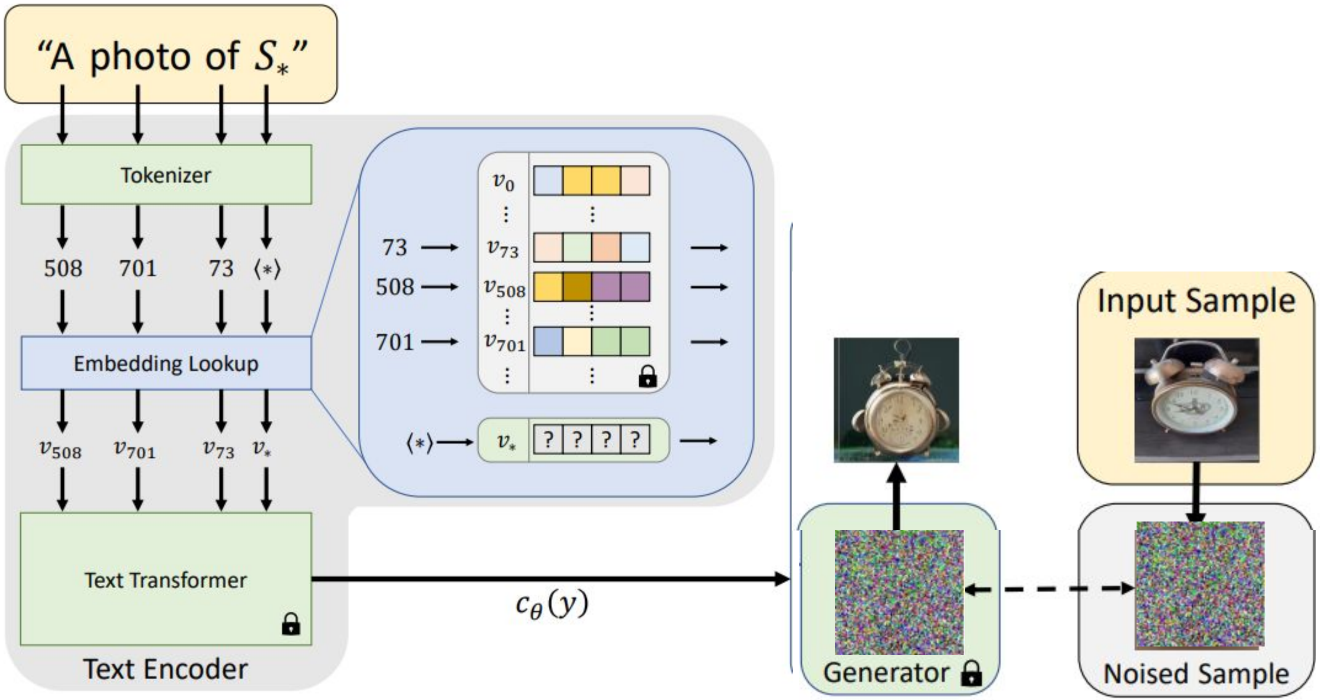
Denoising score matching
[Vincent, 2011]

Ho et al., 2020

$$x_i = \sqrt{\bar{\alpha}_i} x_0 + \sqrt{1 - \bar{\alpha}_i}\epsilon, \;\; \epsilon \sim \mathcal{N}(0, I)$$

$$q_(x_i|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_i} x_0, (1 - \bar{\alpha}_i) I)$$

$$\nabla_x \log q(x_i|x_0) = -\frac{\epsilon}{\sqrt{1-\bar{\alpha}_i}}$$



"A photo of S_*"
Tokenizer
508
701
73
⟨*⟩
Embedding Lookup
v_{508}
v_{701}
v_{73}
v_*
Text Transformer
Text Encoder
v_0
v_{73}
v_{508}
v_{701}
v_*
? ? ? ?
$c_\theta(y)$
Generator
Input Sample
Noised Sample

Figure 3: Object variations generated using our method, the CLIP-based reconstruction of DALLE-2 (Ramesh et al., 2022), and human captions of varying lengths. Our method generates variations which are typically more faithful to the original subject.

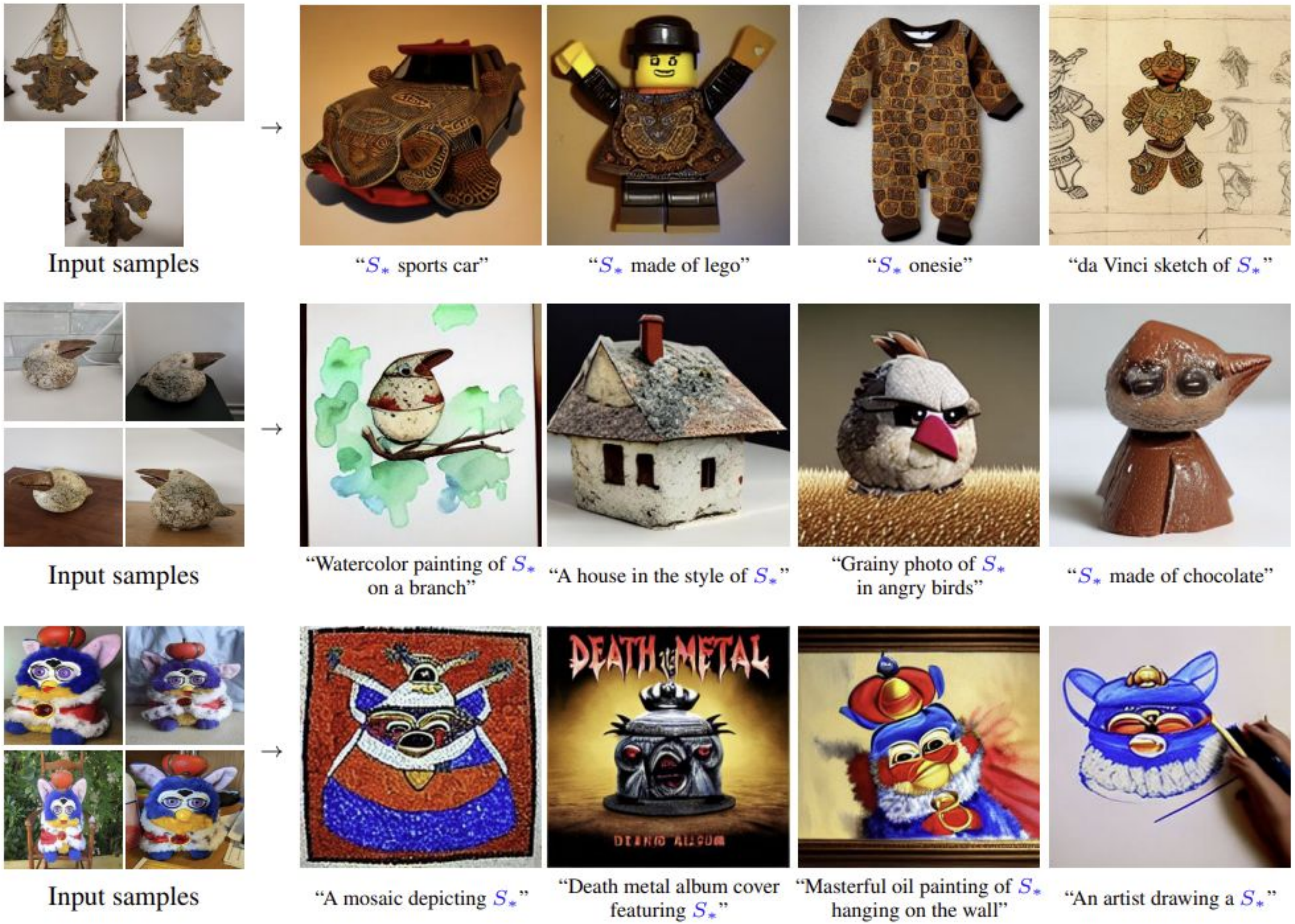

Input samples
"S_* sports car"
"S_* made of lego"
"S_* onesie"
"da Vinci sketch of S_*"
Input samples
"Watercolor painting of S_* on a branch"
"A house in the style of S_*"
"Grainy photo of S_* in angry birds"
"S_* made of chocolate"
Input samples
"A mosaic depicting S_*"
DEATH METAL
"Death metal album cover featuring S_*"
"Masterful oil painting of S_* hanging on the wall"
"An artist drawing a S_*"

Figure 6: The textual-embedding space can represent more abstract concepts, including styles. This allows us to discover words which can be used for style-guided generation. Image credits: @QinniArt (top), @David Revoy (bottom). Image reproduction authorized for non-commercial use only.

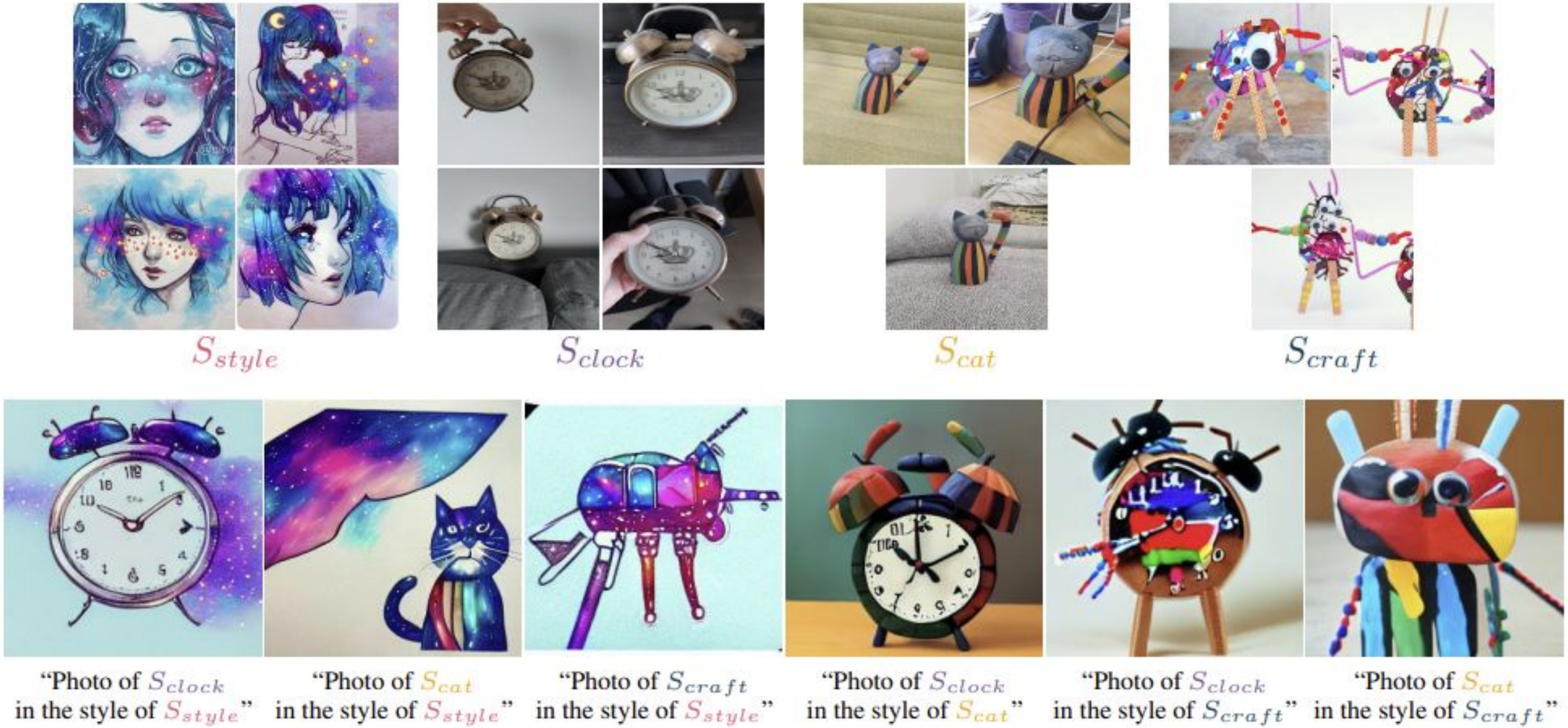


Figure 7: Compositional generation using two learned pseudo-words. The model is able to combine the semantics of two concepts when using a prompt that combines them both. It is limited in its ability to reason over more complex relational prompts, such as placing two concepts side-by-side. Image credits: @QinniArt (left), @Leslie Manlapig (right). Reproductions authorized for non-commercial / non-print use respectively.

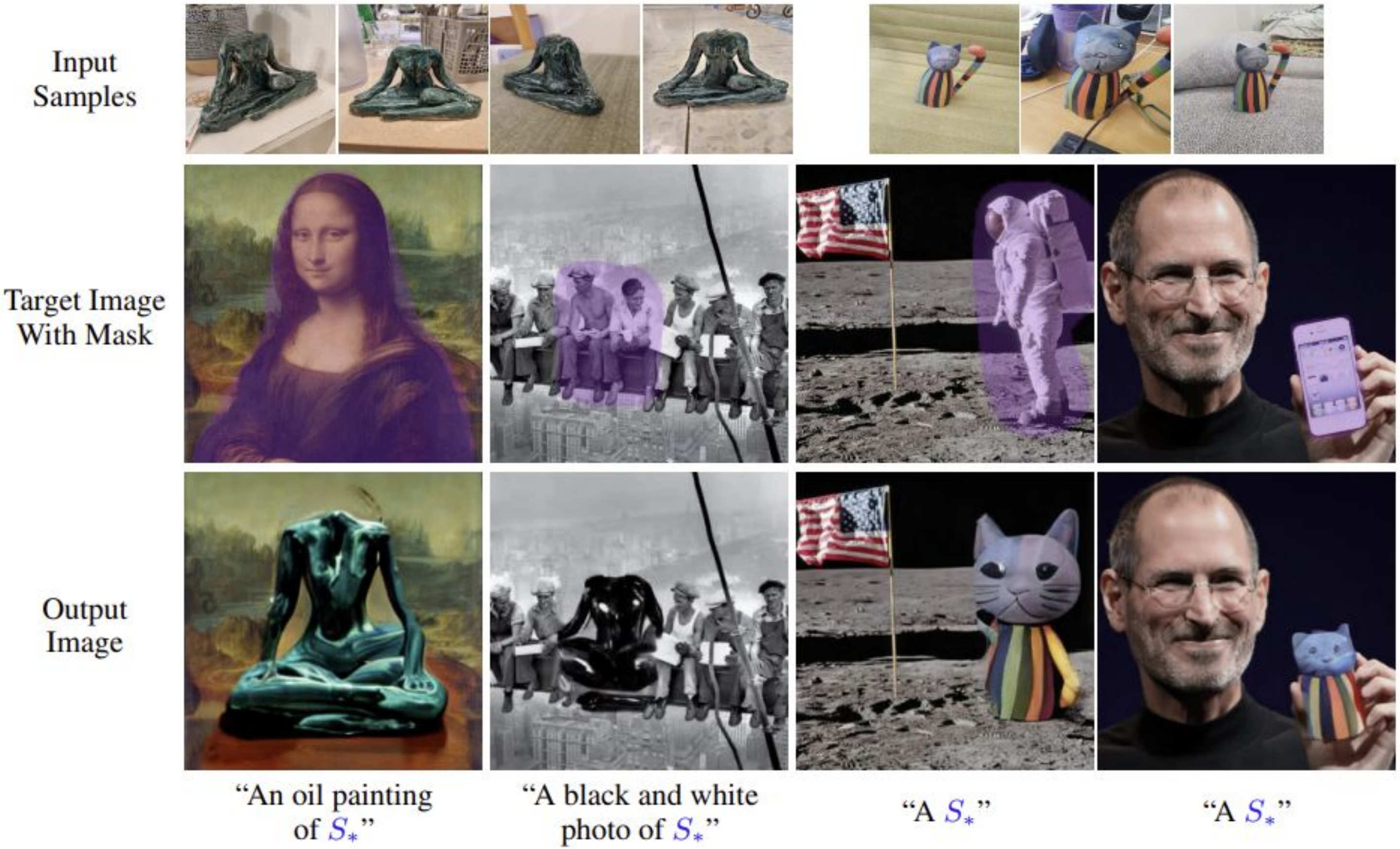


Figure 9: Our words can be used with downstream models that build on LDM. Here, we perform localized image editing using Blended Latent Diffusion (Avrahami et al., 2022a)