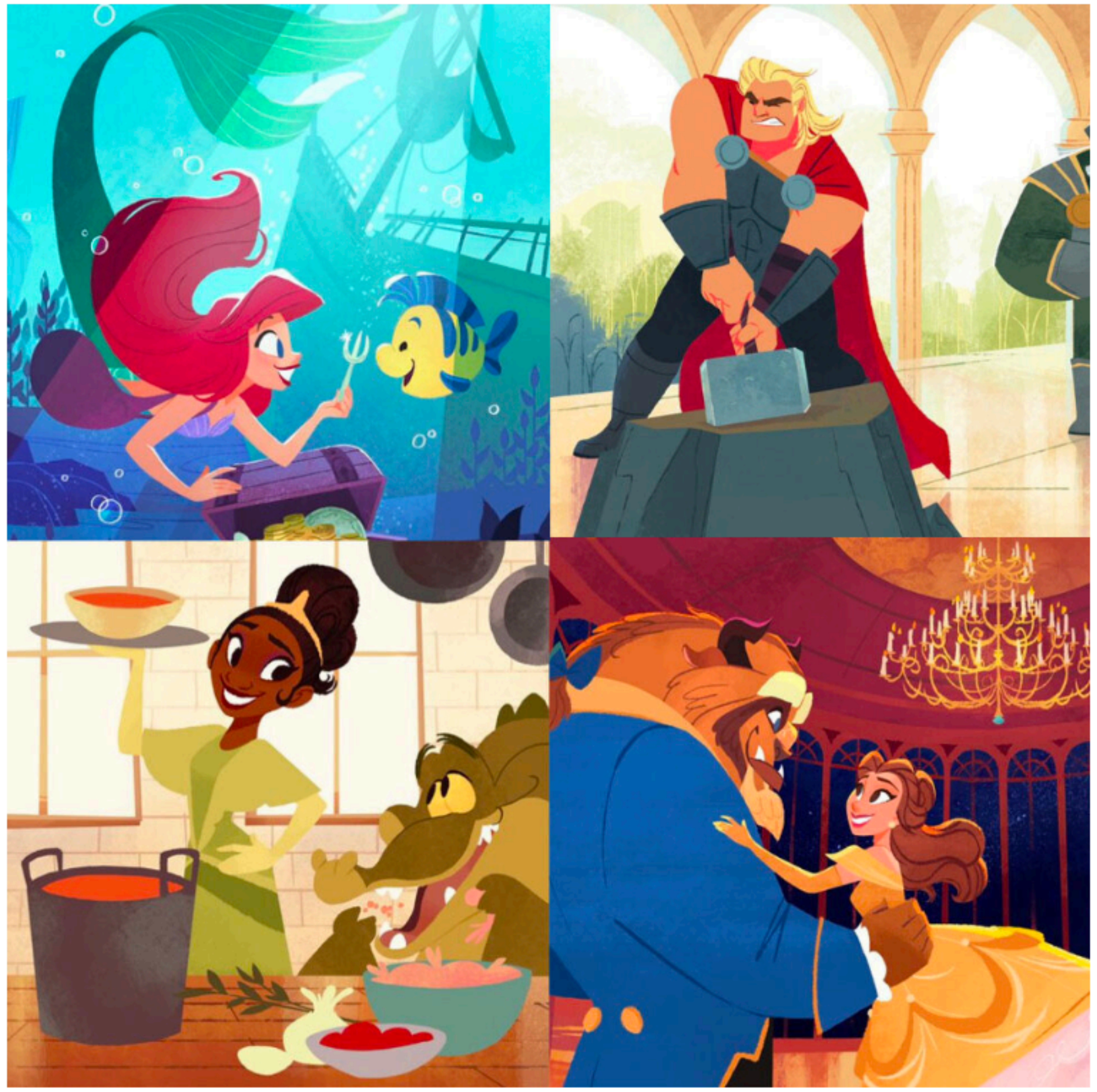
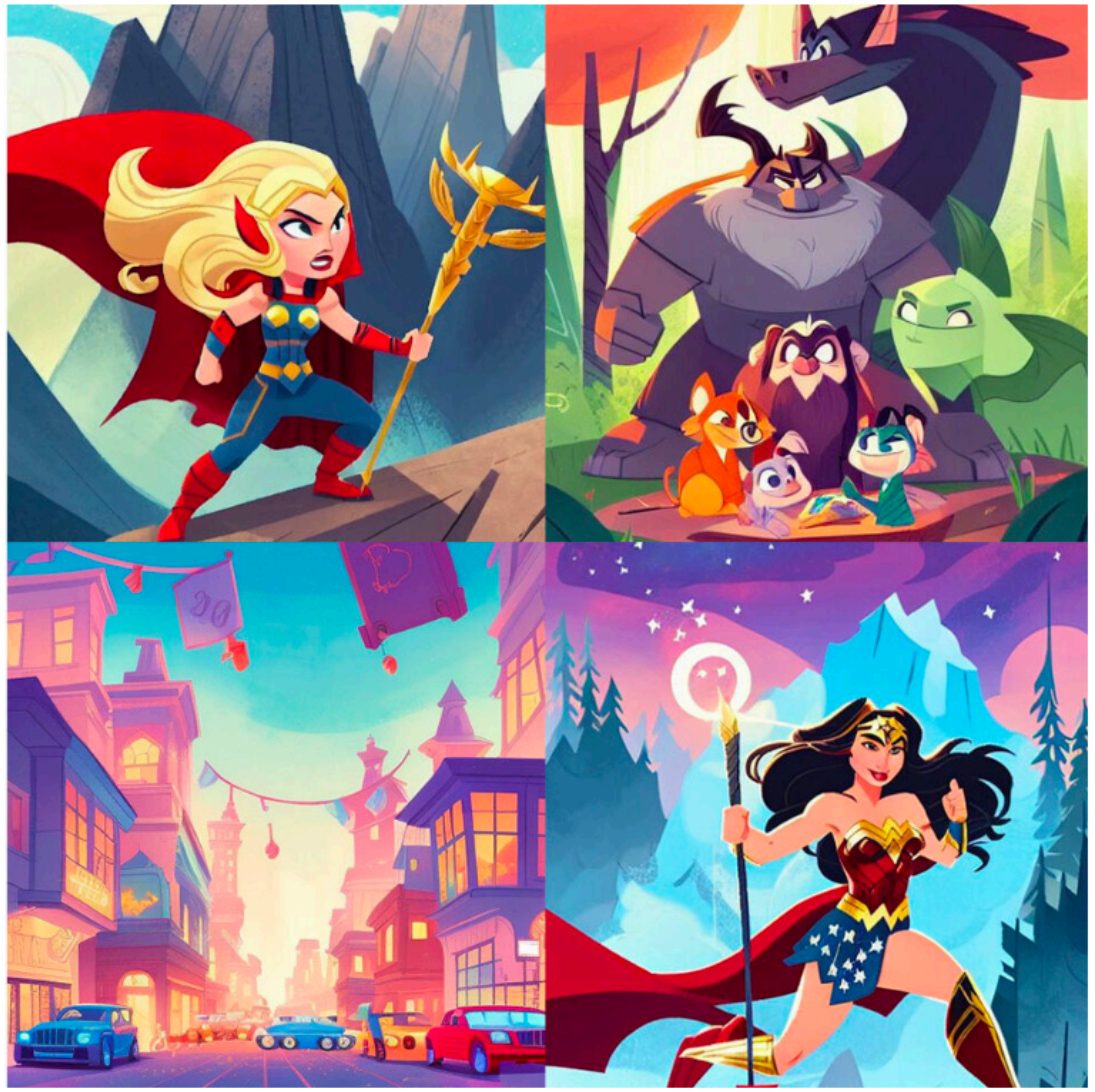
Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models

11/14/2023

AI plagiarizing the style of artist Hollie Mengert

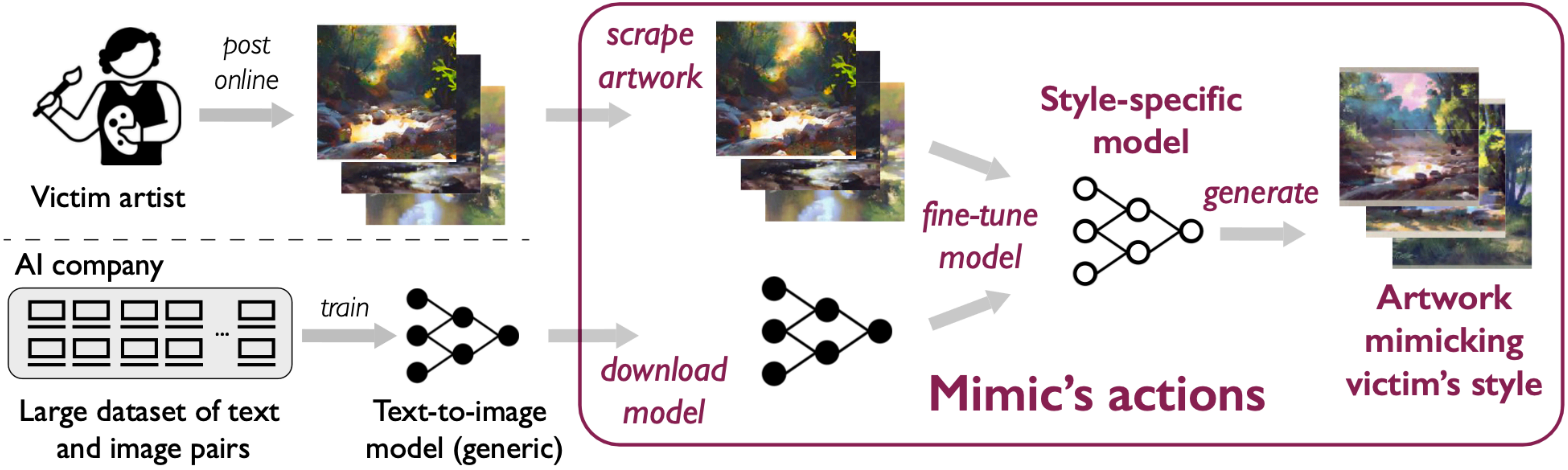


Original artwork
by Hollie Mengert

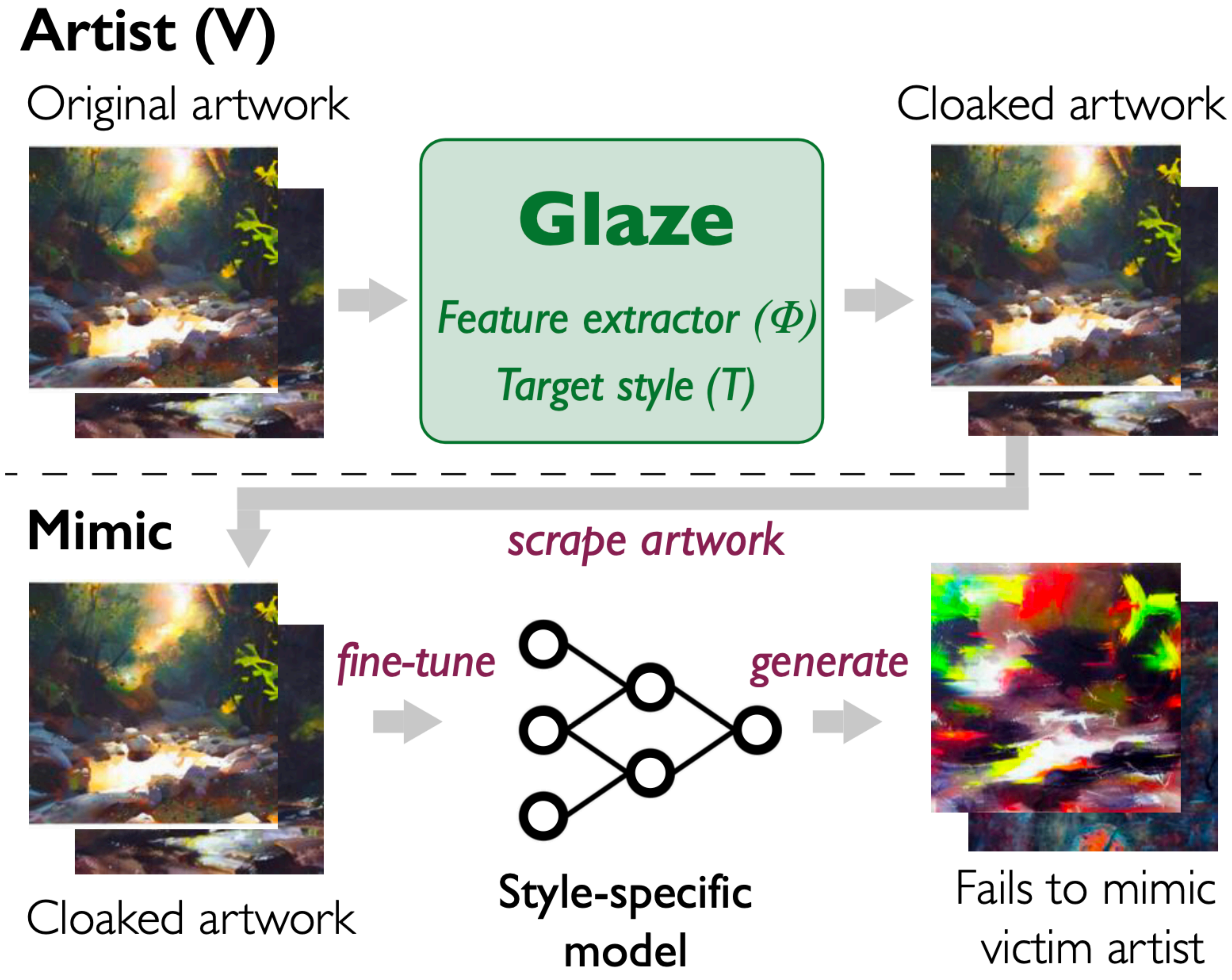


Mimicked artwork
in Hollie's style

Mimicry Attack



Overview of Glaze



Artist (V)

Original artwork

Glaze

Feature extractor (Φ)

Target style (T)

Cloaked artwork

Mimic

scrape artwork

fine-tune

generate

Cloaked artwork

Style-specific model

Fails to mimic victim artist

Text-to-Image Models

Model training

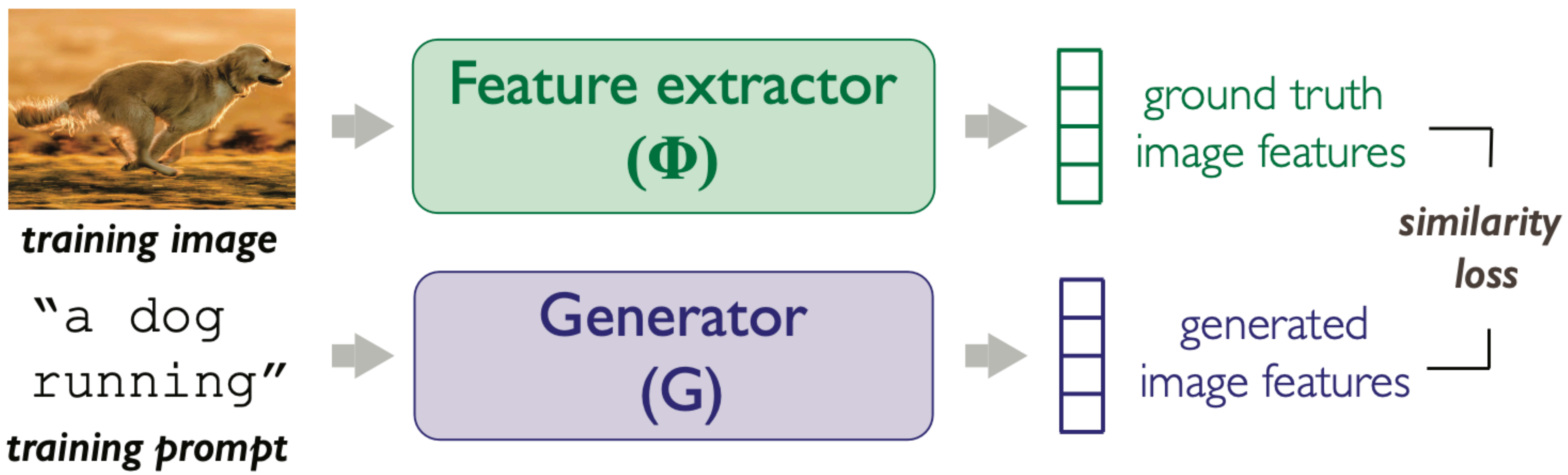
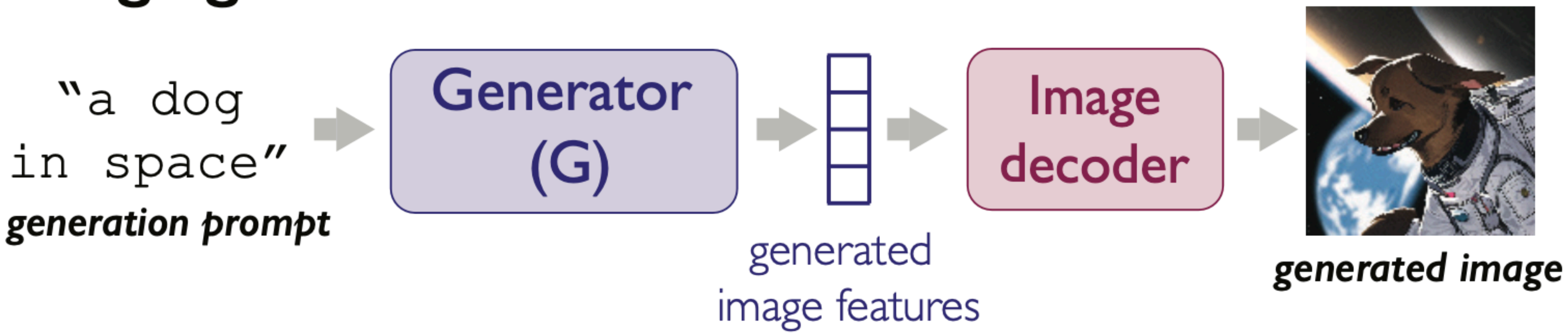
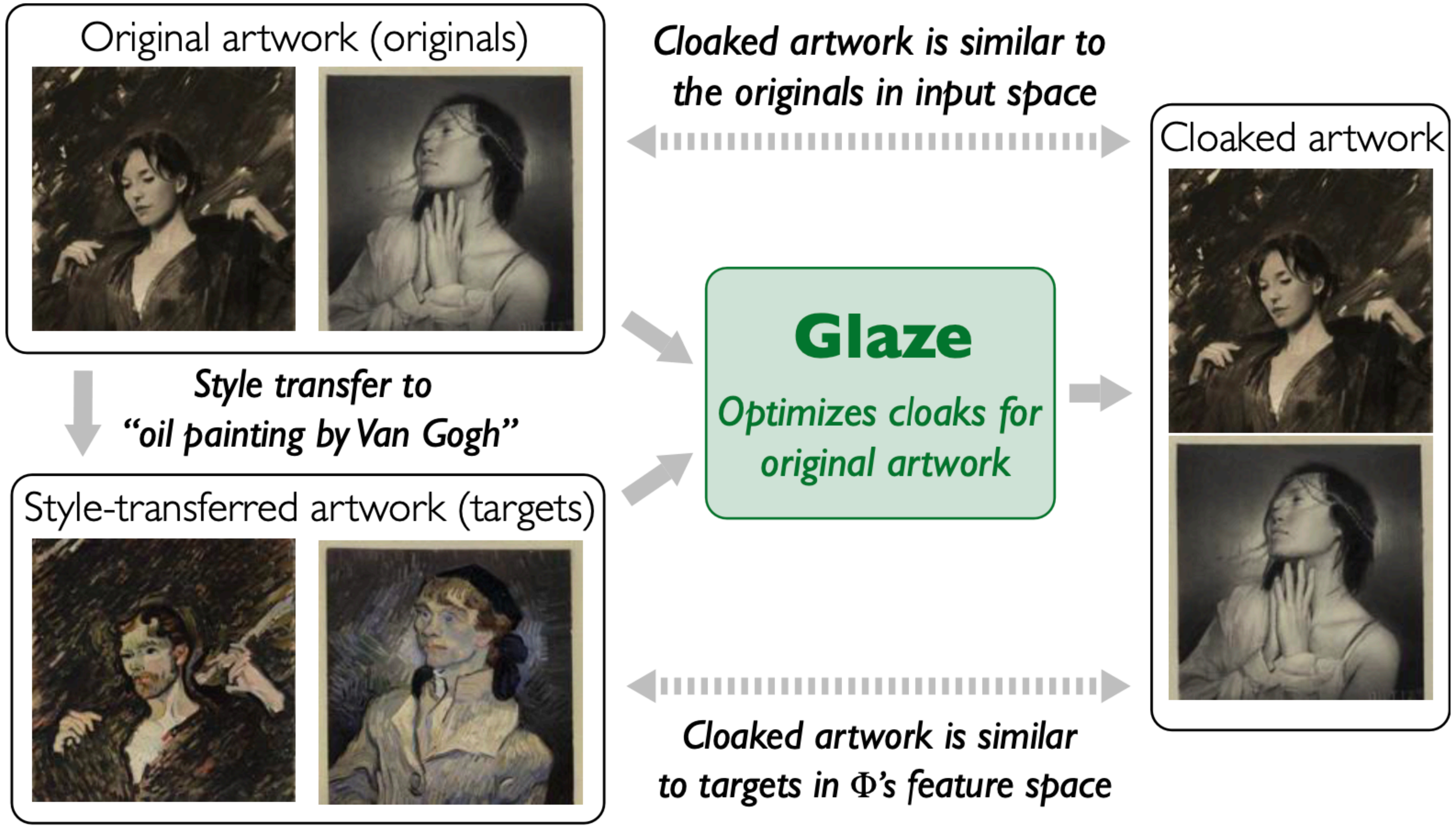


Image generation



Key Idea of Glaze



Original artwork (originals)
Style transfer to
"oil painting by Van Gogh"
Style-transferred artwork (targets)
Cloaked artwork is similar to
the originals in input space
Glaze
Optimizes cloaks for
original artwork
Cloaked artwork
Cloaked artwork is similar
to targets in Φ's feature space

a) Style transfer **b) Cloak optimization**

Objective

$$\min_{\delta_x} Dist\left(\Phi(x+\delta_x), \Phi(\Omega(x,T))\right),$$

$$\text{subject to } \ |\delta_x| < p,$$

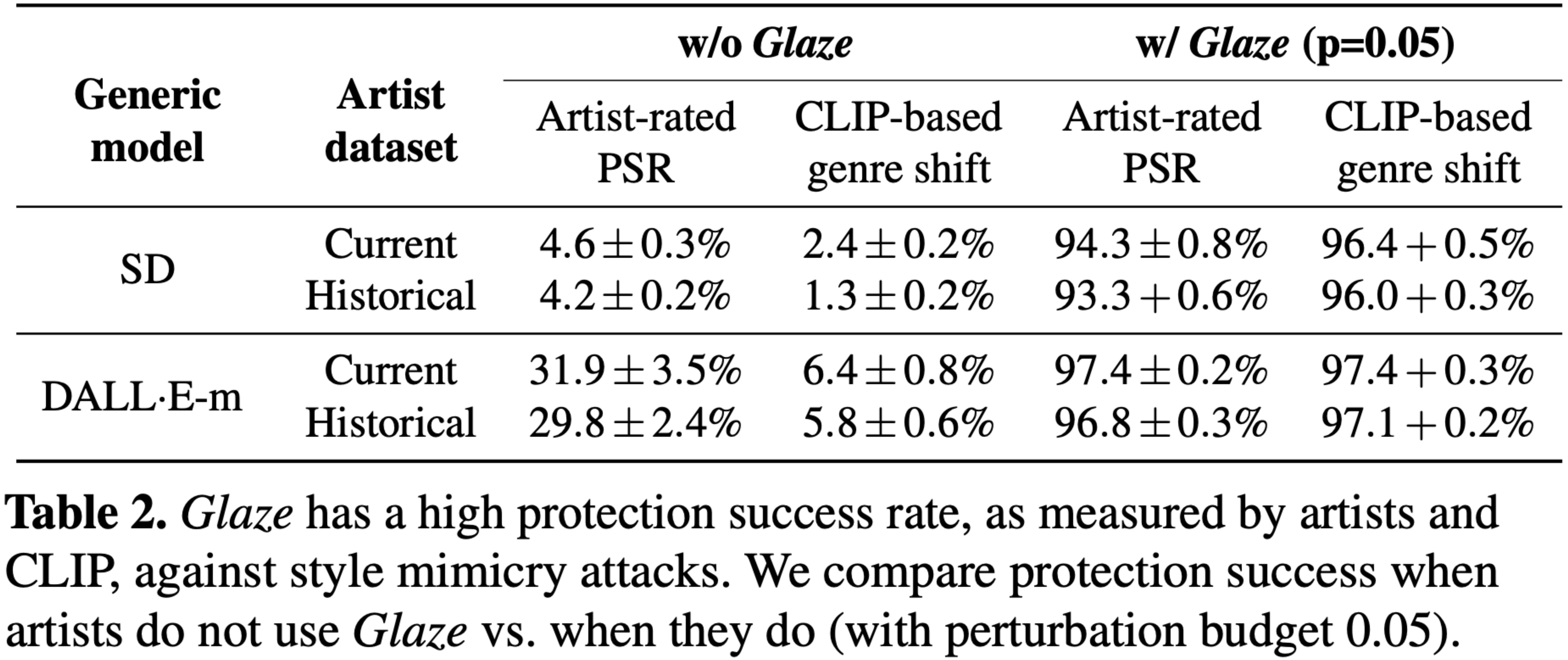
Objective

$$\min_{\delta_x} ||\Phi(\Omega(x,T)), \Phi(x+\delta_x)||_2^2 + \alpha \cdot max(LPIPS(\delta_x) - p, 0)$$

Evaluation

- Artists’ perception
  - 1,156 artist participants
  - The percent of participants who rated Glaze’s protection as “successful” or “very successful.”
- CLIP-based genre shift
  - Classify art images into art genres
  - The percentage of mimicked art whose top 3 predicted genres do not contain the original genre

Main Results



Generic model	Artist dataset	w/o *Glaze*		w/ *Glaze* (p=0.05)	
		Artist-rated PSR	CLIP-based genre shift	Artist-rated PSR	CLIP-based genre shift
SD	Current	4.6 ± 0.3%	2.4 ± 0.2%	94.3 ± 0.8%	96.4 + 0.5%
	Historical	4.2 ± 0.2%	1.3 ± 0.2%	93.3 + 0.6%	96.0 + 0.3%
DALL·E-m	Current	31.9 ± 3.5%	6.4 ± 0.8%	97.4 ± 0.2%	97.4 + 0.3%
	Historical	29.8 ± 2.4%	5.8 ± 0.6%	96.8 ± 0.3%	97.1 + 0.2%

Table 2. *Glaze* has a high protection success rate, as measured by artists and CLIP, against style mimicry attacks. We compare protection success when artists do not use *Glaze* vs. when they do (with perturbation budget 0.05).

Discussions

- Evasion?