Directional Textual Inversion for Personalized Text-to-Image Generation

Kunhee Kim[1*] NaHyeon Park[1*] Kibeom Hong[2] Hyunjung Shim[1]

KAIST[1] SMU[2]

Project page

Text-to-Image Personalization

- Targeted adaptation of models to learn representations of novel, user-provided concepts
- Allows creation of customized images that faithfully render given concept (e.g., face, style)

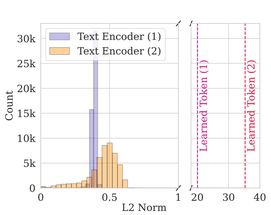
Textual Inversion

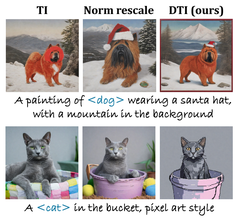
- Text embedding optimization to represent the concept V*
- Limitations in text prompt adherence ☹ *But why?*

Our work

✓ Analyzes why standard TI fails to faithfully express user prompts in the output image
  ✓ Both experimentally, and mathematically
✓ Proposes a simple yet effective solution to this ☺

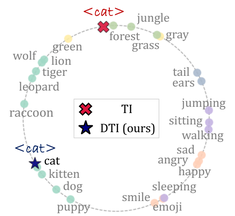
Observation 1: Excessive Norms





Observation 2: Semantic Drift

- Learned token embeddings drift away from semantically related concepts
- Direction no longer aligns with class tokens (e.g., "dog", "cat")
- The embedding loses meaningful semantic structure



Why do excessive norms lead to reduced text fidelity?

Effect 1. Positional information is attenuated

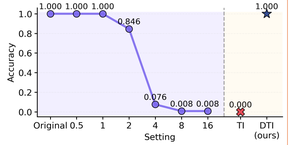
- Large norms suppress aditive positional signals after normalization
- The model becomes less sensitive to token position

⇒ Weaker contextual understanding

Effect 2. Residual updates stagnate

- Residual updates are bounded after normalization
- Large hidden states barely change direction

⇒ Layers behave like near-identity mappings



Method: Directional Textual Inversion

Core Idea

- Decompose text embedding into: $\boldsymbol{e} = m^* \boldsymbol{v}, \|\boldsymbol{v}\| = 1$
- Fix magnitude, learn only the direction

Key Design

- Fix magnitude m^* to in-distribution scale
- Optimize only direction $\boldsymbol{v}$ (semantic carrier)

How We Optimize

- Use *Riemannian SGD* on the sphere
- Add von Mises-Fisher prior = directional regularization

$$\mathcal{L}_{prior}(\boldsymbol{v}) = -\kappa^\top \boldsymbol{v}$$

Why It Works

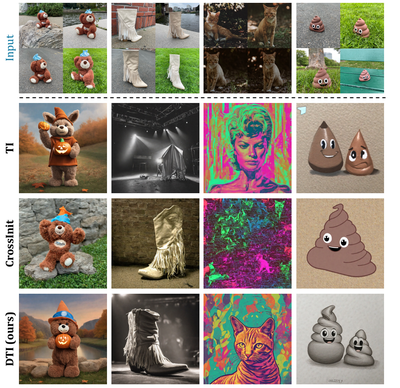
- Prevents norm explosion
- Preserves semantic geometry
- Improves conditioning in Transformer

Experiments

Methods	SDXL		SANA 1.5-1.6B		SANA 1.5-4.8B	
	Image	Text	Image	Text	Image	Text
TI	**0.561**	0.292	**0.480**	0.621	0.446	0.646
TI-rescaled	0.243	0.466	0.253	0.655	0.287	0.548
CrossInit	0.545	0.464	0.344	0.614	0.299	0.622
DTI (ours)	0.450	**0.522**	0.479	**0.744**	**0.452**	**0.757**

	TI	CrossInit	DTI (ours)
Image fidelity	13.78	42.87	**43.45**
Text alignment	10.83	22.40	**66.77**

Optimizer	m^*	$\kappa \times 10^{-3}$	Image	Text
AdamW	mean	0.1	0.335	0.463
RSGD	min	0.1	0.030	0.074
RSGD	5.0 (OOD)	0.1	0.383	0.373
RSGD	mean	0.0	**0.507**	0.436
RSGD	mean	0.5	0.278	**0.688**
RSGD	mean	0.1	0.450	0.522



Takeaways

- **Direction = semantics**
- TI fails due to **norm inflation**
- Large norms:
  - Hide position
  - Block refinement
- **DTI = freeze norm + learn direction**
- Improves **prompt faithfulness** with minimal cost
- Enables **smooth interpolation & creative control**