# Mitigating Semantic Collapsing in Generative Personalization with Test-time Embedding Adjustment

Anh Bui[1], Thuy-Trang Vu[1], Trung Le[1], Junae Kim[2], Tamas Abraham[2], Rollin Omari[2], Amar Kaur[2], Dinh Phung[1]
[1] Monash University, [2] DSTG

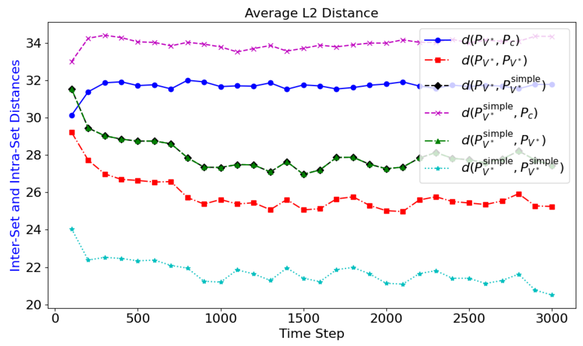
## The "Semantic Collapsing Problem"

Well known phenomenon ⬇ but don't understand Why yet

$$G(\lfloor p, V^* \rfloor) \rightarrow G(V^*)$$

Finding 1: the phenomenon ⬆ because of the SCP ⬇

$$\tau(\lfloor p, V^* \rfloor) \rightarrow \tau(V^*)$$

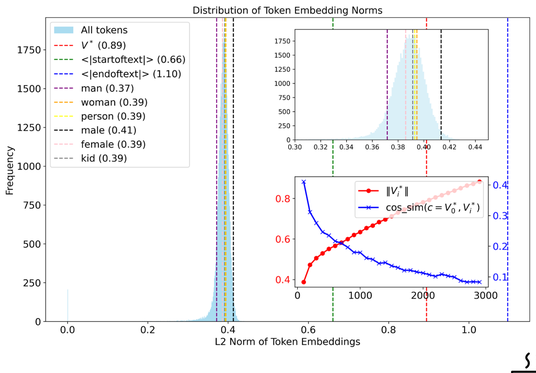



Finding 2: Root cause of SCP because of the unconstrained optimization in personalization process, i.e., TI or DB

TI: $$\min_{v^*} \mathbb{E}_{x,p,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_{t,\epsilon}, t, \lfloor p, V^* \rfloor)\|_2^2\right]$$

DB: $$\min_{\theta,v^*} \mathbb{E}_{x,p,x_{pr},p_{pr},\epsilon,\epsilon',t}\left[\|\epsilon - \epsilon_\theta(x_{t,\epsilon}, t, \lfloor p, V^* \rfloor)\|_2^2 + \lambda\left\|\epsilon' - \epsilon_\theta\left(x_{t',\epsilon'}^{pr}, t', p^{pr}\right)\right\|_2^2\right]$$

Embedding of V* can be arbitrary large and shifted from c



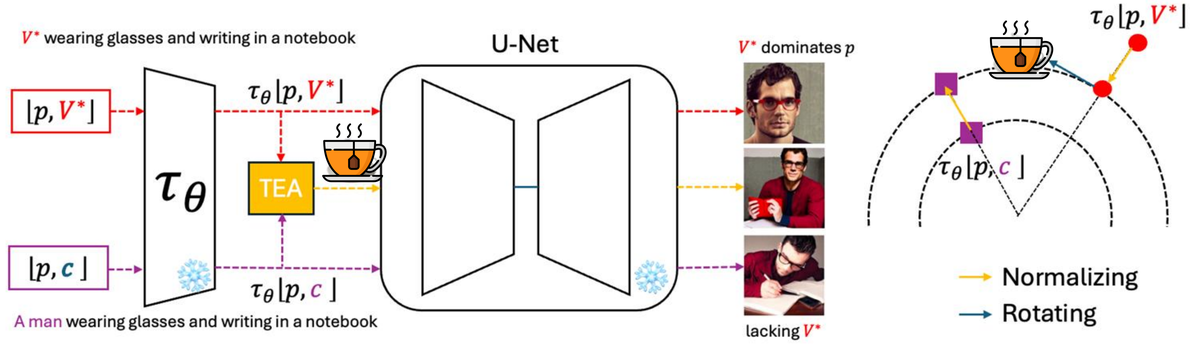

## Test-time Embedding Adjustment

$$\hat{\tau}(\lfloor p, V^* \rfloor) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)}\tilde{\tau}(\lfloor p, V^* \rfloor) + \frac{\sin(\alpha\theta)}{\sin(\theta)}\tilde{\tau}(\lfloor p, c \rfloor)$$

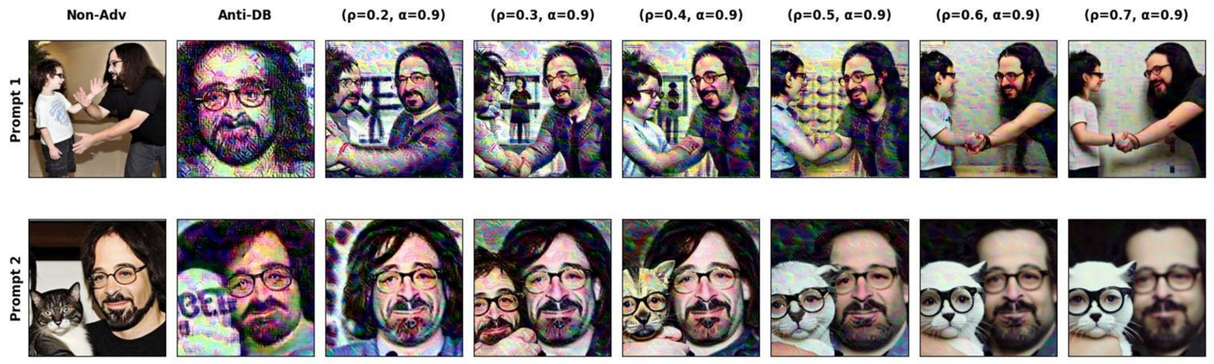
$$\tilde{\tau}(\lfloor p, V^* \rfloor) = \beta\|\tau(\lfloor p, c \rfloor)\|\frac{\tau(\lfloor p, V^* \rfloor)}{\|\tau(\lfloor p, V^* \rfloor)\|}$$

$$\tilde{\tau}(\lfloor p, c \rfloor) = \beta\|\tau(\lfloor p, c \rfloor)\|\frac{\tau(\lfloor p, c \rfloor)}{\|\tau(\lfloor p, c \rfloor)\|}$$

$$\theta = \arccos(\tilde{\tau}(\lfloor p, c \rfloor), \tilde{\tau}(\lfloor p, V^* \rfloor))$$


## Anti-Personalization with




Non-Adv: Standard Personalization →Attackers can generate inappropriate images of anyone
Anti-DB: Anti-Personalization → Defenders protect users' images by adding invisible mask → Attackers fail
Underlying mechanism: Adversarial mask makes the SCP worse → V* over-domination → generating distortion
Anti-DB with TEA → TEA can mitigate SCP → reducing concept domination → generate clean images

## Qualitative Results

OminiControl ⬆ vs OminiControl with ⬇
Prompt: "V* wearing glasses"

## Quantitative Results

| Method | $CLIP_T^p$ ↑ | $CLIP_T^f$ ↑ | CLIP-I ↑ | DINO-I ↑ | VLM-P ↑ | VLM-I ↑ |
|---|---|---|---|---|---|---|
| | | | *CC101 - Pet Dog* | | | |
| ES | 18.54 | 26.02 | 61.33 | 43.71 | 64.25 | 74.00 |
| ES+TEA | 18.72 (+0.18) | 26.11 (+0.09) | 64.56 (+3.23) | 48.32 (+4.61) | 66.50 (+2.25) | 77.25 (+3.25) |
| | | | *CC101 - Plushie Teddybear* | | | |
| ES | 20.48 | 26.80 | 81.64 | 49.08 | 78.00 | 80.25 |
| ES+TEA | 20.61 (+0.13) | 27.3 (+0.50) | 82.84 (+1.20) | 51.17 (+2.09) | 80.25 (+2.25) | 81.50 (+1.25) |
| | | | *Subject - Clock* | | | |
| OC | 18.11 | 23.90 | 81.37 | 32.41 | 67.50 | 62.25 |
| OC+TEA | 18.78 (+0.67) | 23.98 (+0.08) | 83.10 (+1.73) | 34.48 (+2.07) | 71.75 (+4.25) | 64.50 (+2.25) |
| | | | *Subject - Oranges* | | | |
| OC | 21.49 | 27.62 | 70.43 | 30.33 | 68.50 | 53.00 |
| OC+TEA | 21.60 (+0.11) | 27.70 (+0.08) | 71.90 (+1.47) | 31.64 (+1.31) | 70.00 (+1.50) | 55.50 (+2.50) |
| | | | *Subject - Penguin* | | | |
| OC | 20.30 | 31.61 | 78.58 | 45.59 | 86.25 | 83.25 |
| OC+TEA | 20.33 (+0.03) | 32.02 (+0.41) | 80.64 (+2.06) | 49.37 (+3.78) | 90.50 (+4.25) | 86.75 (+3.50) |
| | | | *Relationship - A <Carved by> B* | | | |
| RV | 25.64 | 27.74 | N/A | N/A | N/A | N/A |
| RV+TEA | 27.84 (+2.20) | 30.17 (+2.43) | N/A | N/A | N/A | N/A |
| | | | *Relationship - A <Inside> B* | | | |
| RV | 24.97 | 27.87 | N/A | N/A | N/A | N/A |
| RV+TEA | 25.15 (+0.18) | 28.40 (+0.53) | N/A | N/A | N/A | N/A |
| | | | *Relationship - A <Painted on> B* | | | |
| RV | 23.98 | 30.07 | N/A | N/A | N/A | N/A |
| RV+TEA | 24.38 (+0.40) | 30.35 (+0.28) | N/A | N/A | N/A | N/A |

MORE INSIGHTFUL FINDINGS
tuananhbui89.github.io