蚂蚁集团
ANT GROUP

ACCORD: *A*lleviating *C*oncept *Co*upling through Dependence *R*egularization for Text-to-Image *D*iffusion Personalization

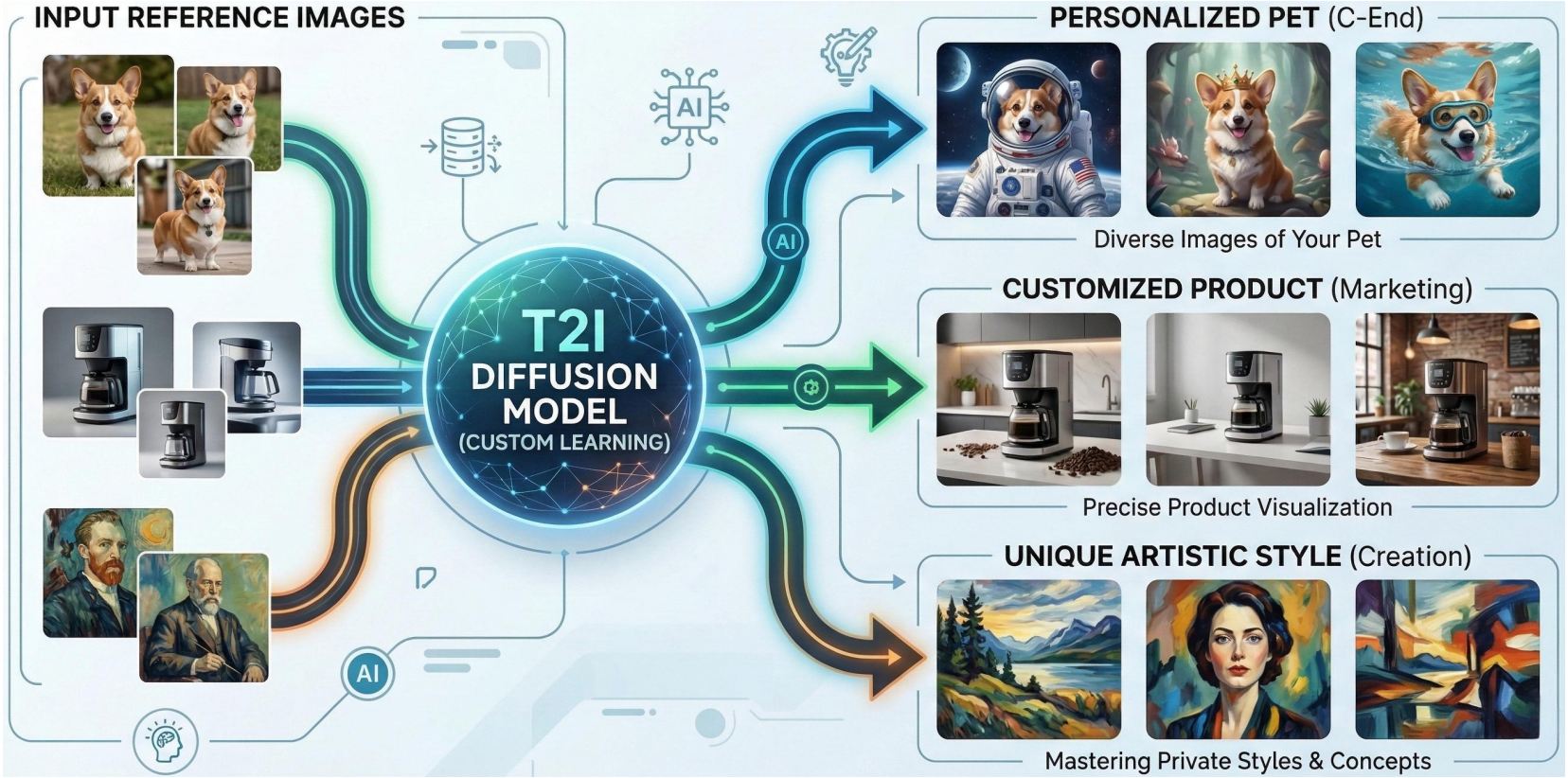
Shizhan Liu, Hao Zheng, Hang Yu, Jianguo Li

Ant Group

ICLR

Customized T2I Diffusion Models: Personalized Generation

- Personalized Generation: Personal Objects | Commercial Posters | Style Transfer



Central Problem: Concept Coupling

- Few-shot data with low variance causes target-context coupling

backpack

girl

Prior Dependence

$$\frac{p(backpack, girl)}{p(backpack)p(girl)}$$

$\ll$

Conditional Dependence

$$E_{x\sim X_{\theta 1}}\left[\frac{p(backpack^*, girl|x)}{p(backpack^*|x)p(girl|x)}\right]$$

Reference Images of a personal backpack
Denoted as backpack*

Finetuning

Finetuning w/ concept decoupling

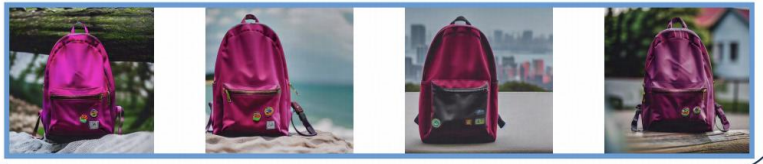
Text-to-Image Model

Model θ_1 Inference

Model θ_2 Inference

Generated Images
Sampled from Generation Distribution $X_{\theta 1}, X_{\theta 2}$

A backpack on top of pink fabric	A backpack on the beach	A backpack with a city in the background	A backpack with a blue house in the background



Motivation

- Existing approaches mitigate overfitting, ACCORD directly formalizes and minimizes concept coupling

Method Category	Representative Methods	Limitations regarding Concept Coupling
Data Regularization	DreamBooth, CustomDiffusion	Use superclass datasets to preserve model priors but risks distorting concept relationships
Weight Regularization	LoRA, SVDiff	Constrains parameter updates to prevent overfitting, which can indiscriminately degrade fidelity
Region/Loss Proxies	CoRe, Facechain-SuDe	Relies on spatial attention heuristics or fails for global attributes like style
Dependence Regularization	**ACCORD (Ours)**	**Directly treats statistical dependencies. Plug-and-play formulation**

Formulation of Concept Coupling

- Measuring concept dependence via **conditional dependence coefficient** r:
  - c_p: Personalization target
  - c_g: General concepts, e.g. text condition
  - $x_{\theta,t}$: Image generated at diffusion timestep t

$$r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t}) = \frac{p(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t})}{p(\mathbf{c}_p|\mathbf{x}_{\theta,t})p(\mathbf{c}_g|\mathbf{x}_{\theta,t})}$$

  - $r(c_p, c_g|x_{\theta,t}) = 1$: the two concepts c_p and c_g are independent
  - $r(c_p, c_g|x_{\theta,t}) \gg 1$ or $r(c_p, c_g|x_{\theta,t}) \ll 1$: the two concepts are strongly correlated
- Concept coupling can be formulated as an excessive inter-concept dependency:

$$\mathbb{E}_{\mathbf{x}_\theta}[|\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,0}) - \log r(\mathbf{c}_s, \mathbf{c}_g)|] \gg 0$$

where c_g denotes the superclass of c_p. Namely the generated image $x_{\theta,0}$ introduce **additional dependencies** between the personalization target c_p and general concepts c_g

Formulation of Concept Coupling

- Target of concept decoupling:
  - Correct $r(c_p, c_g|x_{\theta,0})$ in the generated images so that it approximates the prior concept dependence between c_s and c_g.

$$\mathbb{E}_{\mathbf{x}_\theta}[|\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,0}) - \log r(\mathbf{c}_s, \mathbf{c}_g)|] \to 0$$

- The dependence discrepancy can be further decomposed:

Theorem 1. *The dependence discrepancy can be decomposed into the following two terms:*

$$\mathbb{E}_{\mathbf{x}_\theta}\Big[|\underbrace{\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,0}) - \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_T)}_{\text{① Denoising Dependence Discrepancy}} + \underbrace{\log r(\mathbf{c}_p, \mathbf{c}_g) - \log r(\mathbf{c}_s, \mathbf{c}_g)}_{\text{② Prior Dependence Discrepancy}}|\Big],$$

where $\mathbf{x}_T$ denotes multivariate standard Gaussian noise.

  - $log\ r(c_p, c_g|x_T) = log\ r(c_p, c_g)$ holds since x_T is Gaussian noise sampled independently of the condition c_p and c_g.

- We then minimize this two discrepancies respectively.

Denoising Decouple Loss (DDLoss)

- Minimize the denoising dependence discrepancy
  - Relax it with the sum of dependence discrepancies between adjacent denoising steps:

$$
\begin{aligned}
|\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,0}) - \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_T)| &= |\sum_{t=1}^{T} \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t-1}) - \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t})| \\
&\leq \sum_{t=1}^{T} |\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t-1}) - \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t})|.
\end{aligned}
$$

  - By Bayes' theorem and the Gaussianity of noisy latents at timestep $t-1$:

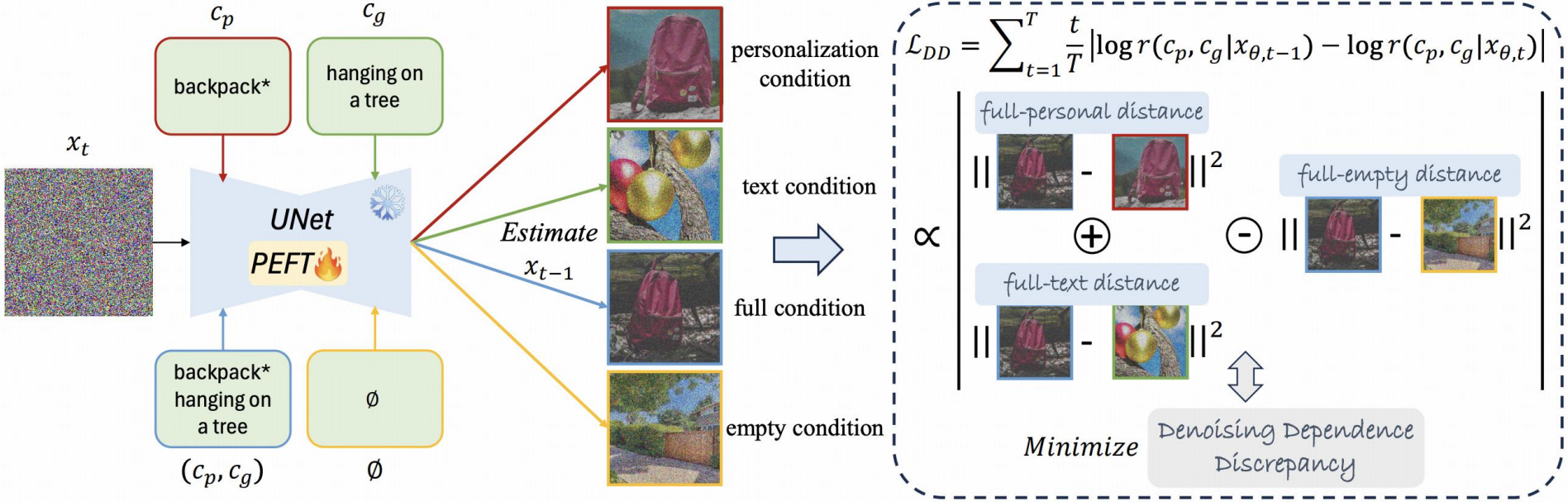
Theorem 2. *The dependence discrepancy between successive time steps in diffusion models can be computed as:*

$$
\begin{aligned}
&\log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t-1}) - \log r(\mathbf{c}_p, \mathbf{c}_g|\mathbf{x}_{\theta,t}) \\
&= \frac{1}{2\boldsymbol{\sigma}_t^2}\Big[\|\mathcal{U}_\theta(\mathbf{x}_t, (\mathbf{c}_p, \mathbf{c}_g), t) - \mathcal{U}_\theta(\mathbf{x}_{\theta,t}, \mathbf{c}_p, t)\|^2 + \|\mathcal{U}_\theta(\mathbf{x}_t, (\mathbf{c}_p, \mathbf{c}_g), t) - \mathcal{U}_\theta(\mathbf{x}_{\theta,t}, \mathbf{c}_g, t)\|^2 \\
&\quad - \|\mathcal{U}_\theta(\mathbf{x}_t, (\mathbf{c}_p, \mathbf{c}_g), t) - \mathcal{U}_\theta(\mathbf{x}_{\theta,t}, \varnothing, t)\|^2\Big],
\end{aligned}
$$

where $\varnothing$ denotes an empty control condition.

Denoising Decouple Loss (DDLoss)

- Calculation of DDLoss:



Prior Decouple Loss (PDLoss)

- Minimize the prior dependence discrepancy
  - Rewrite it as:

$$\log r(\mathbf{c}_p, \mathbf{c}_g) - \log r(\mathbf{c}_s, \mathbf{c}_g) = \log \frac{p(\mathbf{c}_g|\mathbf{c}_p)}{p(\mathbf{c}_g|\mathbf{c}_s)}$$

  - PDLoss estimation relies on conditional probabilities $p(c_g|c_p)$ and $p(c_g|c_s)$, which CLIP can help:

Lemma 2. *For an observation $\mathbf{c}_j$ and condition $\mathbf{c}_k$, the InfoNCE objective seeks to estimate a function $\mathcal{F}(\mathbf{c}_j, \mathbf{c}_k)$ which is proportional to the following density ratio: $\mathcal{F}(\mathbf{c}_j, \mathbf{c}_k) \propto \frac{p(\mathbf{c}_j|\mathbf{c}_k)}{p(\mathbf{c}_j)}$.*

  - By Lemma 2, we have the following approximation:

$$\tau \cos(\mathbf{f}_j, \mathbf{f}_k) \propto \frac{p(\mathbf{c}_j|\mathbf{c}_k)}{p(\mathbf{c}_j)}$$

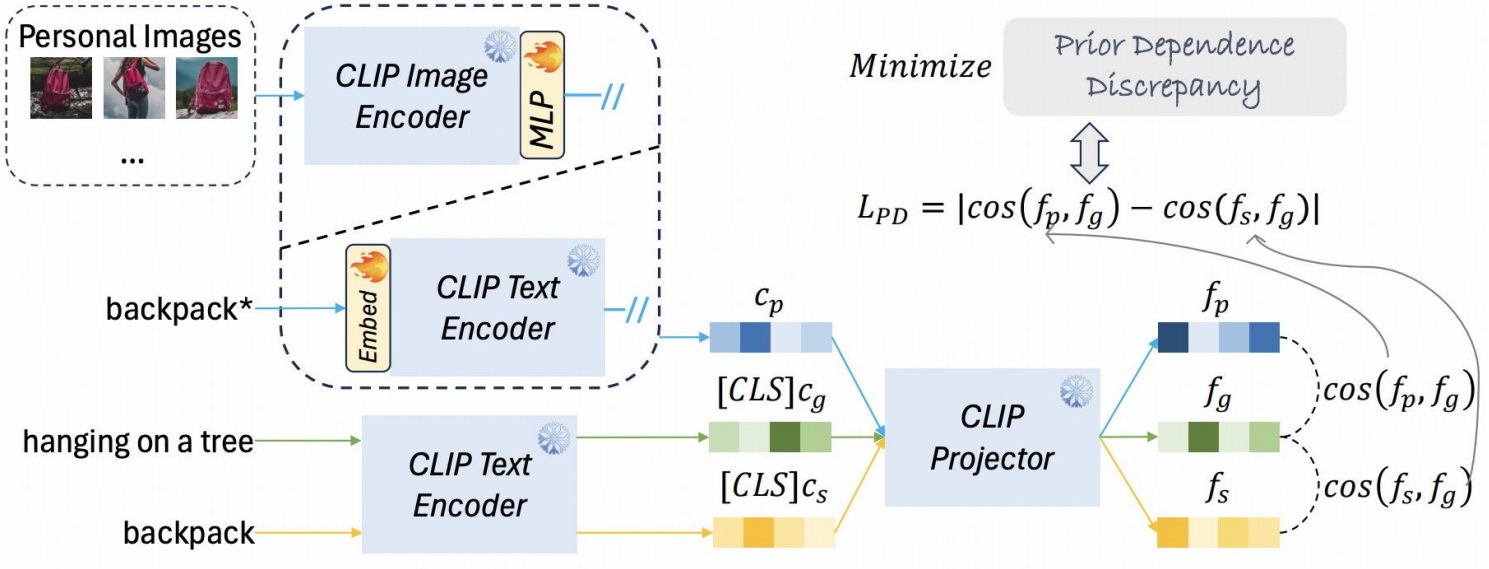
  - Thus, we have Theorem 3:

Theorem 3. *The prior dependence discrepancy can be minimized by the following PDLoss:*

$$\mathcal{L}_{PD} = \mathbb{E}_{\mathbf{c}_g}[|\cos(\mathbf{f}_p, \mathbf{f}_g) - \cos(\mathbf{f}_s, \mathbf{f}_g)|]$$
$$\propto \mathbb{E}_{\mathbf{c}_g}[\left|\frac{p(\mathbf{c}_g|\mathbf{c}_p) - p(\mathbf{c}_g|\mathbf{c}_s)}{p(\mathbf{c}_g)}\right|].$$

Prior Decouple Loss (PDLoss)

- Calculation of PDLoss:



Quantitative Results

- Subject & Style & Face personalization
  - plug-and-play
  - enhancing both text alignment and subject fidelity

Table 1: Quantitative results on DreamBench. The "*" indicates results using per-subject/style loss weights, tuned on a small validation set. "Params." indicates the number of tunable parameters. The W(in)/L(oss) rate is calculated by pairwise human comparison between the anonymous generated results of the baseline and Ours*, with ties omitted. 'PA' denotes percent agreement, namely the percentage of samples receiving consistent judgments from human annotators. The comparison methods improved based on the baseline are *italicized*.

Method	CLIP-T↑	BLIP-T↑	CLIP-I↑	DINO-I↑	W↑/L↓ (%)	PA (%)	Params.
DreamBooth (DB)	30.3	40.3	74.0	69.3	18.1/**75.7**	73.3	819.7 M
CoRe-SD1.5	29.4	40.3	78.3	72.3	19.2/**61.7**	60.0	819.7M
Facechain-SuDe	**31.4**	41.6	74.3	70.5	14.2/**69.2**	70.0	819.7 M
DB w/ Ours	31.1 (+0.8)	42.1 (+1.8)	77.8 (+3.8)	73.5 (+4.2)	-/-	-	819.7 M
DB w/ Ours*	31.3 (+1.0)	**42.1** (+1.8)	**78.6** (+4.6)	**74.4** (+5.1)	-/-	-	819.7 M
CustomDiffusion (CD)	34.2	45.4	62.7	56.9	8.1/**88.1**	76.7	18.3 M
ClassDiffusion	**34.3**	45.8	61.3	55.0	7.5/**75.8**	80.0	18.3M
CD w/ Ours	33.9 (-0.3)	46.4 (+1.0)	71.1 (+8.4)	65.2 (+8.3)	-/-	-	18.3 M
CD w/ Ours*	34.1 (-0.1)	**46.6** (+1.2)	**71.4** (+8.7)	**65.6** (+8.7)	-/-	-	18.3 M
LoRA (SDXL)	34.5	47.0	76.3	72.1	17.6/**70.5**	70.0	92.9 M
SVDiff	32.7	43.7	72.6	66.6	1.7/**85.0**	83.3	0.2 M
Omnigen	**35.3**	47.8	73.9	68.6	30.8/**48.3**	46.7	3.8 B
LoRA w/ Ours	35.1 (+0.6)	**47.8** (+0.8)	76.8 (+0.5)	71.9 (-0.2)	-/-	-	92.9 M
LoRA w/ Ours*	35.2 (+0.7)	47.7 (+0.7)	**77.1** (+0.8)	**72.4** (+0.3)	-/-	-	92.9 M
VisualEncoder (VE)	25.9	36.1	79.1	75.5	21.1/**67.6**	56.7	3.0 M
VE w/ Ours	25.9 (+0.0)	35.8 (-0.3)	80.0 (+0.9)	76.0 (+0.5)	-/-	-	3.0 M
VE w/ Ours*	**26.3** (+0.4)	**36.1** (+0.0)	**80.4** (+1.3)	**76.7** (+1.2)	-/-	-	3.0 M

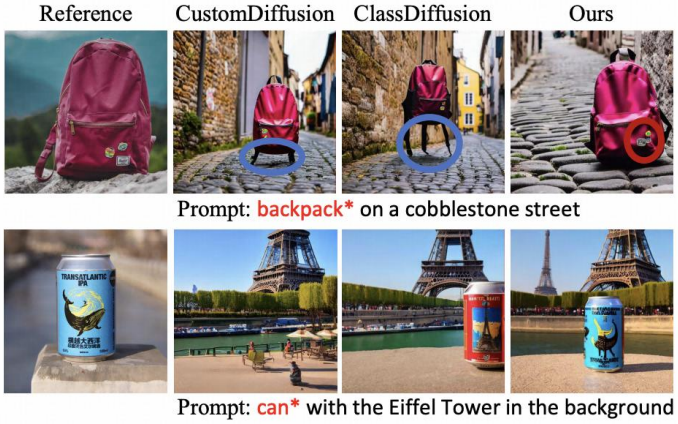
Table 2: Quantitative results on StyleBench. The "*" denotes adjusting DDLoss and PDLoss weights across different styles. "Gram-D" is the gram matrix distance.

Method	CLIP-T↑	BLIP-T↑	Gram-D↓
DreamBooth	31.3	46.6	42728
Facechain-SuDe	31.0	45.8	**39978**
DB w/ Ours	31.9 (+0.6)	**47.3** (+0.7)	42524 (-0.5%)
DB w/ Ours*	**32.0** (+0.7)	47.2 (+0.6)	41911 (-1.9%)
CustomDiffusion	31.2	47.7	53347
ClassDiffusion	31.8	48.4	52998
CD w/ Ours	31.7 (+0.5)	48.5 (+0.8)	48649 (-8.8%)
CD w/ Ours*	**31.8** (+0.6)	**48.5** (+0.8)	**47852** (-10.3%)
LoRA (SDXL)	33.1	49.7	47193
Omnigen	31.9	47.5	45067
B-LoRA	33.0	49.0	**42048**
LoRA (SDXL) w/ Ours	33.6 (+0.5)	50.7 (+1.0)	47693 (+1.1%)
LoRA (SDXL) w/ Ours*	**33.6** (+0.5)	**50.7** (+1.0)	46361 (-1.8%)
VisualEncoder	17.7	30.2	32176
VE w/ Ours	17.7 (+0.0)	30.3 (+0.1)	31382 (-2.5%)
VE w/ Ours*	**18.4** (+0.7)	**30.9** (+0.7)	**27984** (-13.0%)

Table 5: Quantitative results on FFHQ.

Method	CLIP-T↑	BLIP-T↑	Face-Sim↑
IP-Adapter	20.0	34.7	14.8
+ Ours	**20.7** (+0.7)	**34.8** (+0.1)	**16.4** (+1.6)

Qualitative Results



Qualitative Results

Thanks for attention!