Paper

Code & Models

Zero-shot Generative Model Adaptation via Image-specific Prompt Learning

CVPR 2023 WED-AM-311

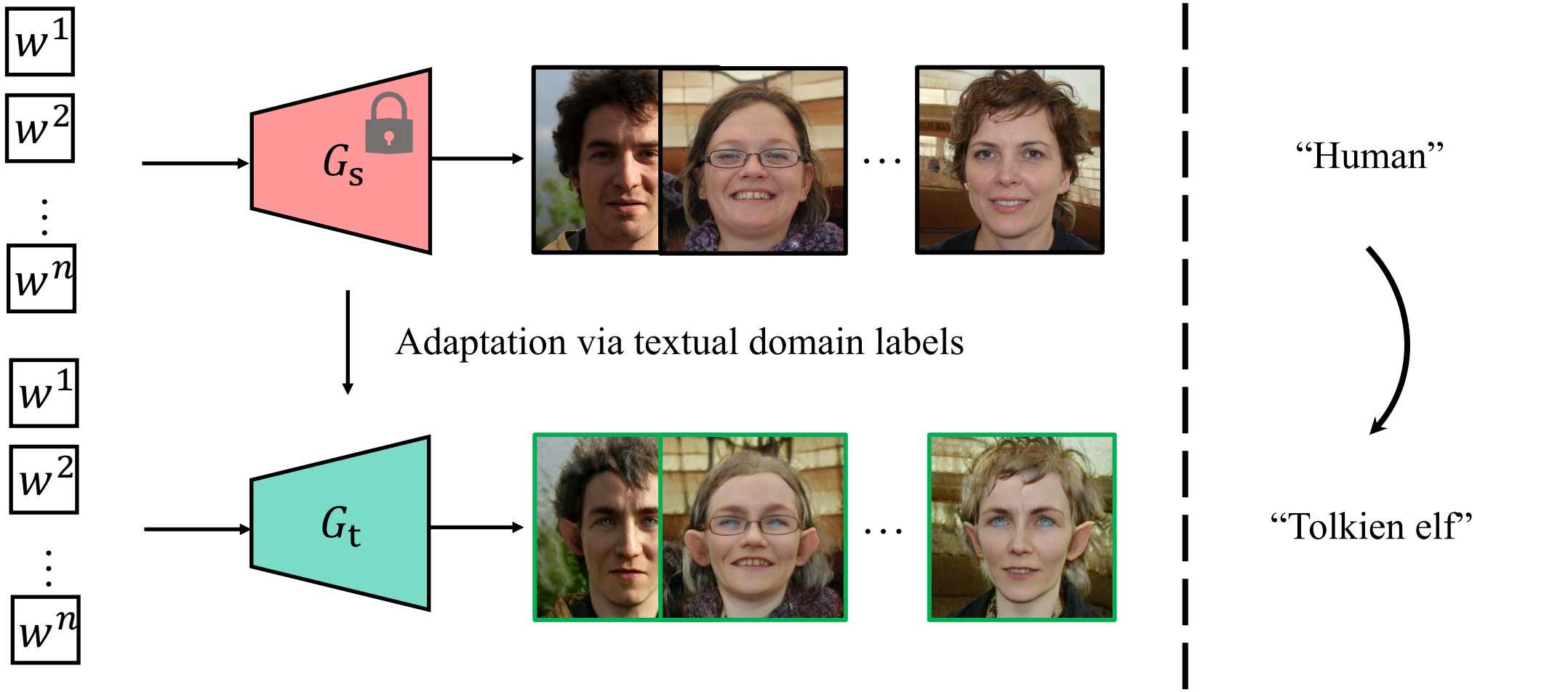
Jiayi Guo[1*], Chaofei Wang[1*], You Wu[2], Eric Zhang[3], Kai Wang[3], Xingqian Xu[3], Shiji Song[1], Humphrey Shi[3,4], Gao Huang[1]

[1]Tsinghua University, BNRist. [2]UCAS. [3]SHI Labs @ Oregon & UIUC. [4]Picsart AI Research.

[*]Equal Contribution

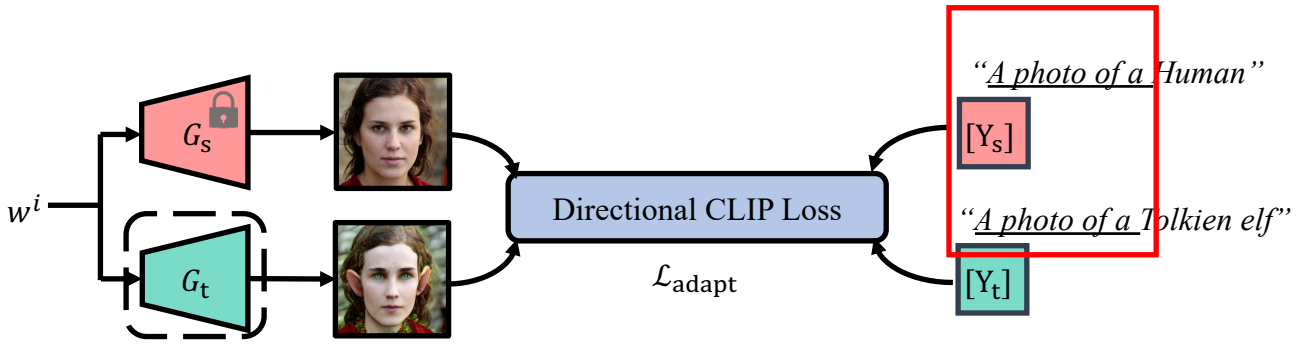
Overview – *Task Definition*

Task definition: Zero-shot generative model adaptation

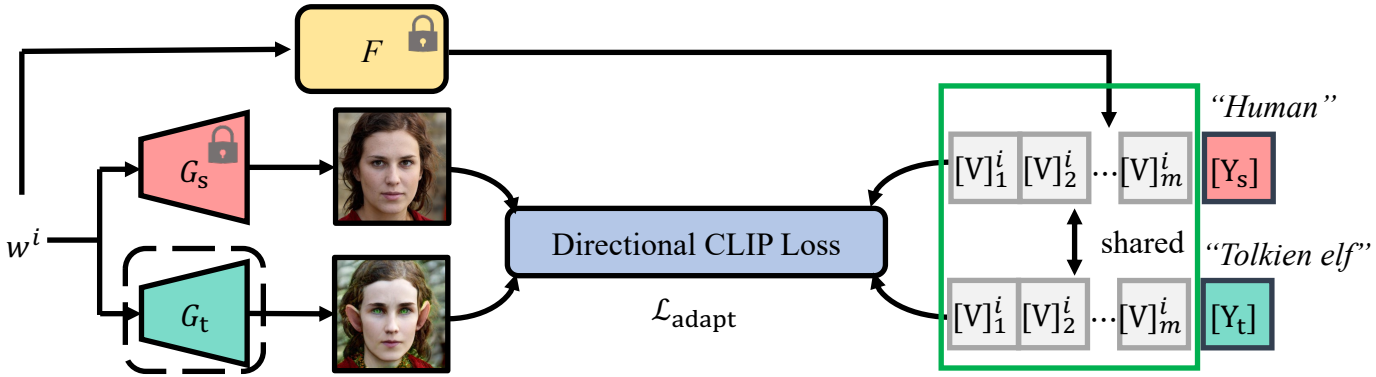


Overview – *Motivation and Contribution*

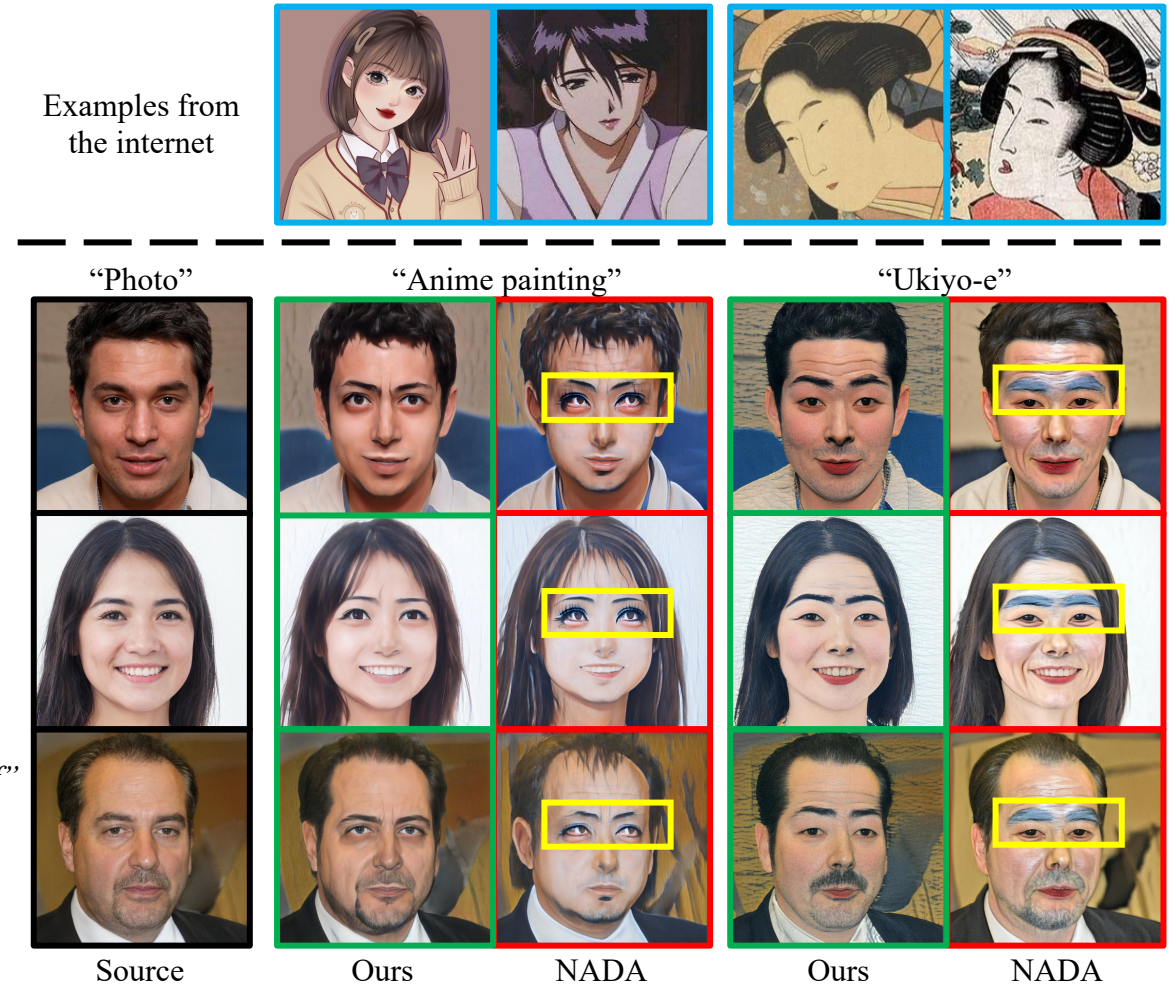
Alleviate the mode collapse issues in existing works



A fixed and shared adaptation direction[1-2]



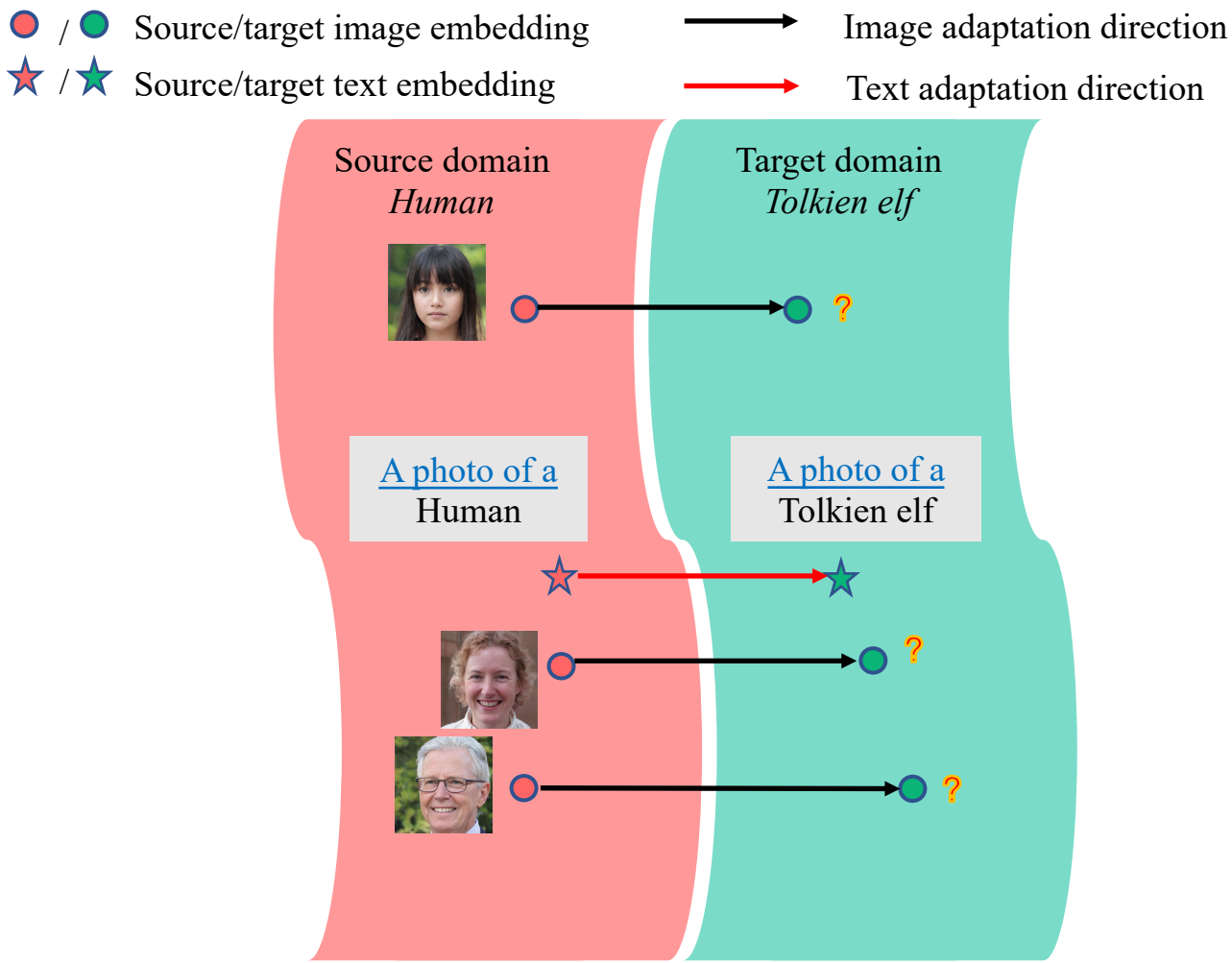
More precise and diversified adaptation directions



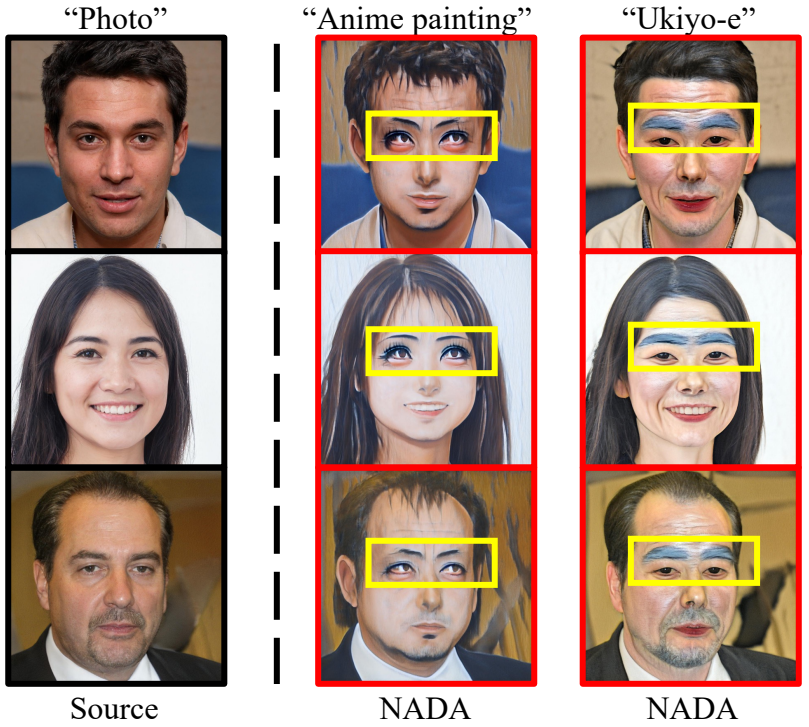
~~***Similar and unexpected patterns***~~

[1] Gal, Rinon, et al. "StyleGAN-NADA: CLIP-guided domain adaptation of image generators." *TOG* 2022.
[2] Kim, Gwanghyun, et al. "DiffusionCLIP: Text-guided diffusion models for robust image manipulation." *CVPR* 2022.

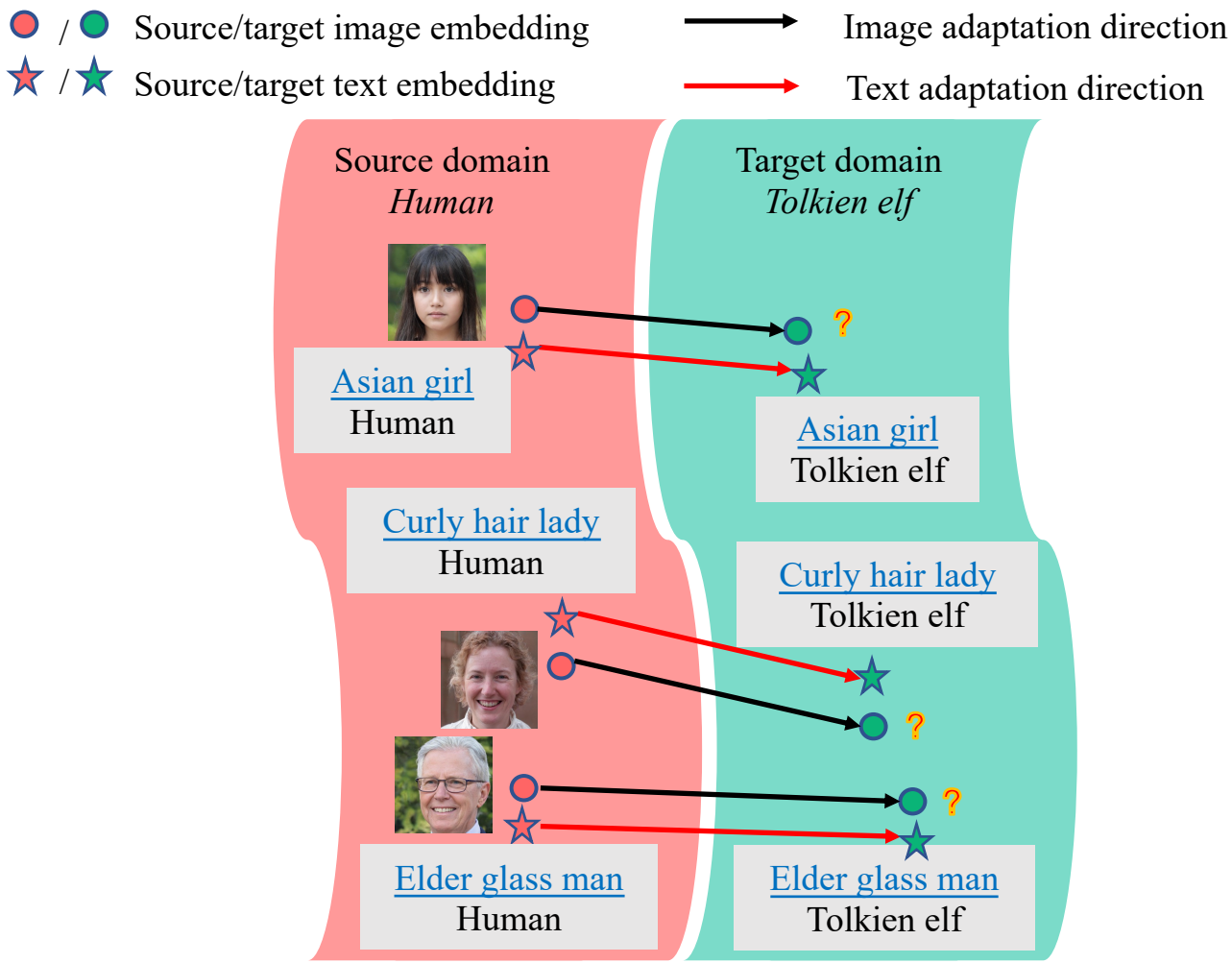
Motivation – *Reason behind the Mode Collapse Issues*



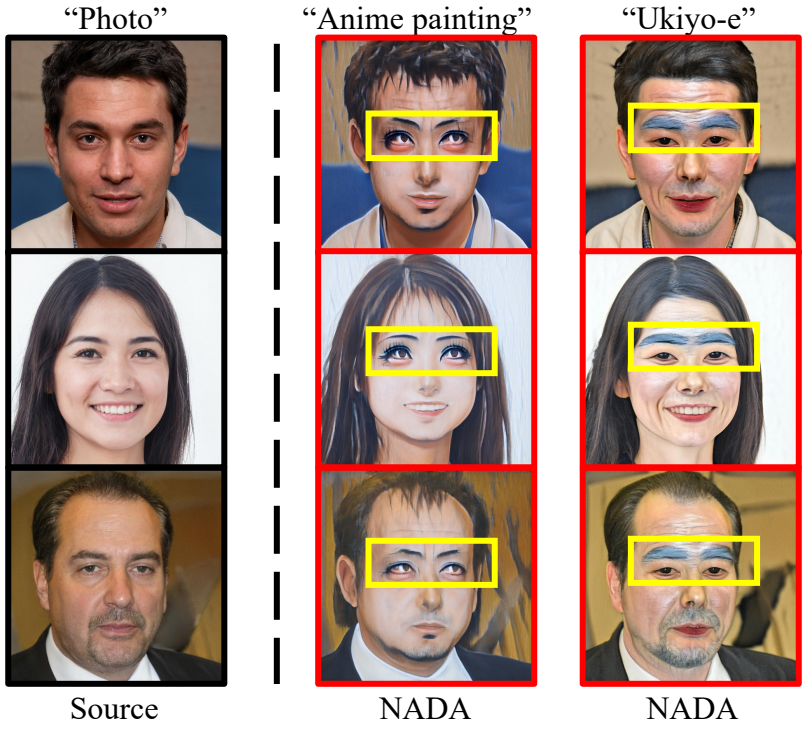
(a) Manual prompts:
a shared fixed reference direction



Motivation – *Reason behind the Mode Collapse Issues*



(b) Learnable prompts (Ours):
image-specific reference directions

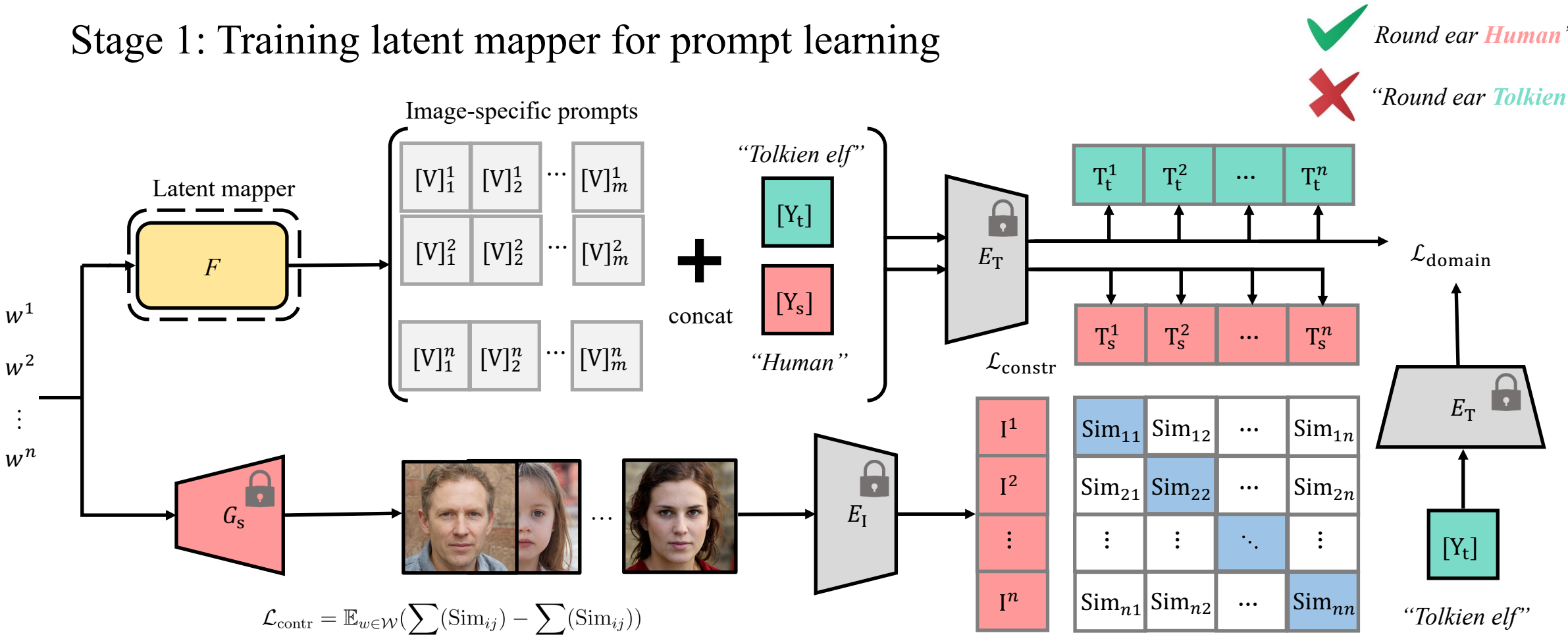


Question:

How to automatically obtain image-specific prompts and reference adaptation directions?

Method – *Image specific Prompt Leaning (IPL)*

Stage 1: Training latent mapper for prompt learning



$$\mathcal{L}_{\text{contr}} = \mathbb{E}_{w \in \mathcal{W}}\left(\sum_{i \neq j}(\text{Sim}_{ij}) - \sum_{i=j}(\text{Sim}_{ij})\right)$$

$$\mathcal{L}_{\text{domain}} = -\mathbb{E}_{w^i \in \mathcal{W}} \sum_{i=1}^{n}\left(\text{Cos}(E_{\text{T}}(\text{M}_{\text{t}}^i), E_{\text{T}}(\text{Y}_{\text{t}}))\right)$$

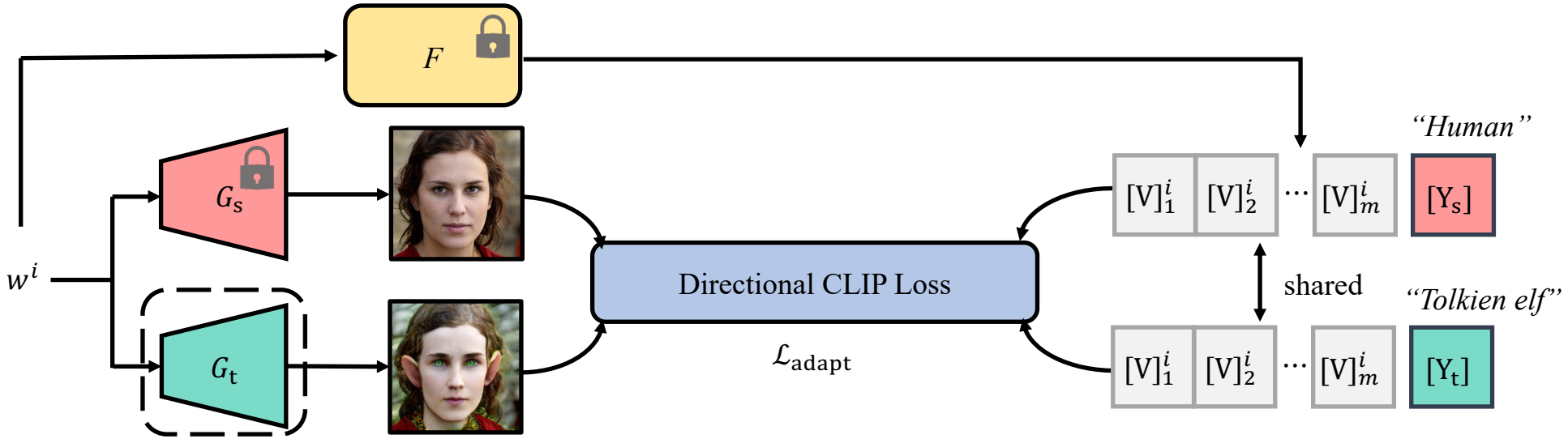
$$\mathcal{L} = \mathcal{L}_{\text{constr}} + \lambda \mathcal{L}_{\text{domain}}$$

Contrastive learning + Domain regularization

Encode image-specific features ***Avoid feature conflict***

Method – *Image specific Prompt Leaning (IPL)*

Stage 2: Training generator for image synthesis

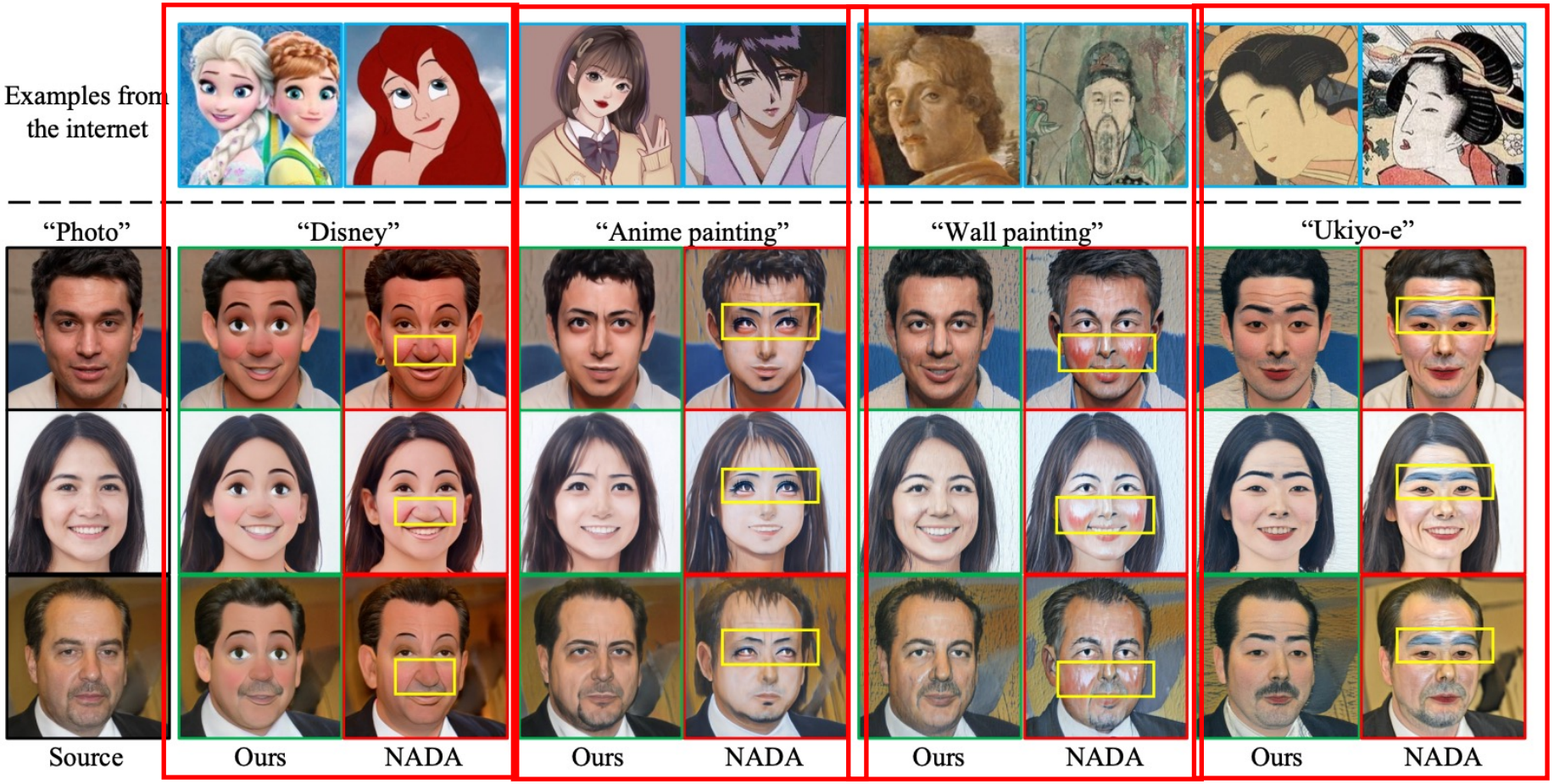


$$\mathcal{L}_{\text{adapt}} = \mathbb{E}_{w^i \in \mathcal{W}} \sum_{i=1}^{n} (1 - \frac{\Delta \text{I}_i \cdot \Delta \text{T}_i}{|\Delta \text{I}_i||\Delta \text{T}_i|})$$

Improved Directional CLIP loss

Experiments – *Generative Model Adaptation*

Qualitative comparison

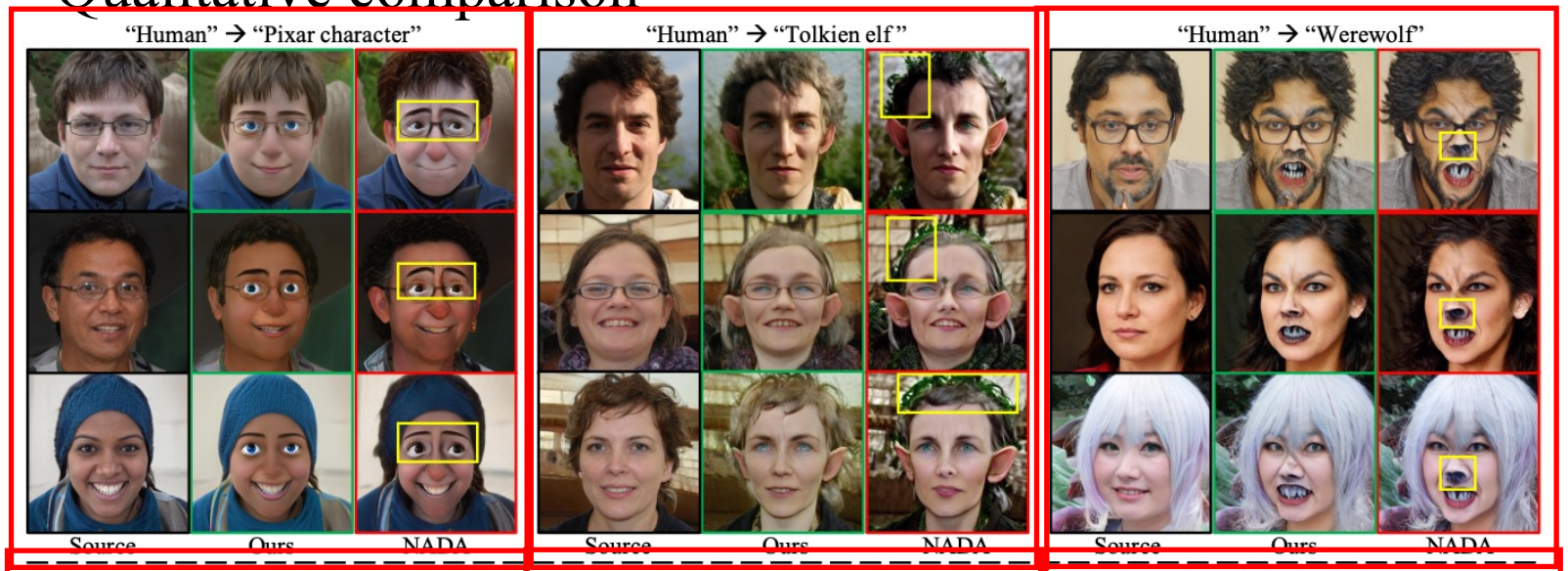


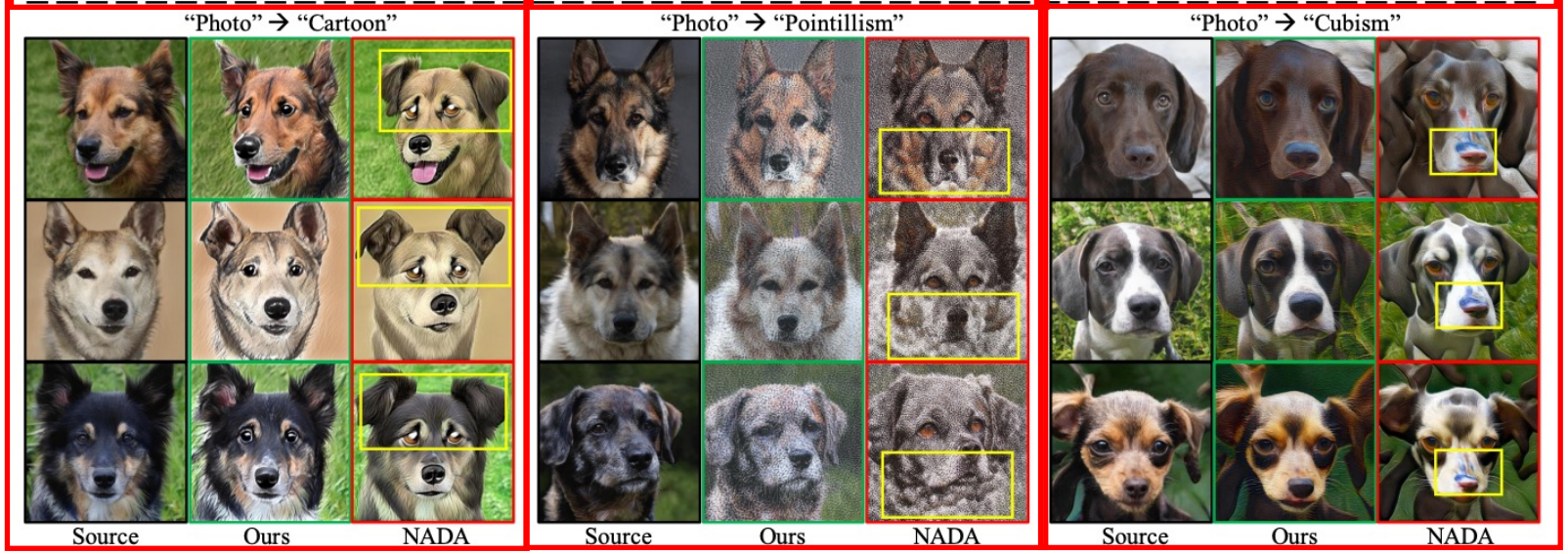
IPL alleviates the mode collapse issues of nasolabial folds (Disney), squinting eyes (Anime painting), red cheeks (Wall painting), and blue eyebrows (Ukiyo-e).

[1] Gal, Rinon, et al. "StyleGAN-NADA: CLIP-guided domain adaptation of image generators." *TOG* 2022.

Experiments – *Generative Model Adaptation*

Qualitative comparison





Depressed emotions (Pixar character),
Green mussy noise on hairs (Tolkien elf),
Ruined noses (Werewolf),
……

Folded ears (Cartoon),
Unshaped necks (Pointillism),
Blue noses (Cubism),
……

Experiments – *Generative Model Adaptation*

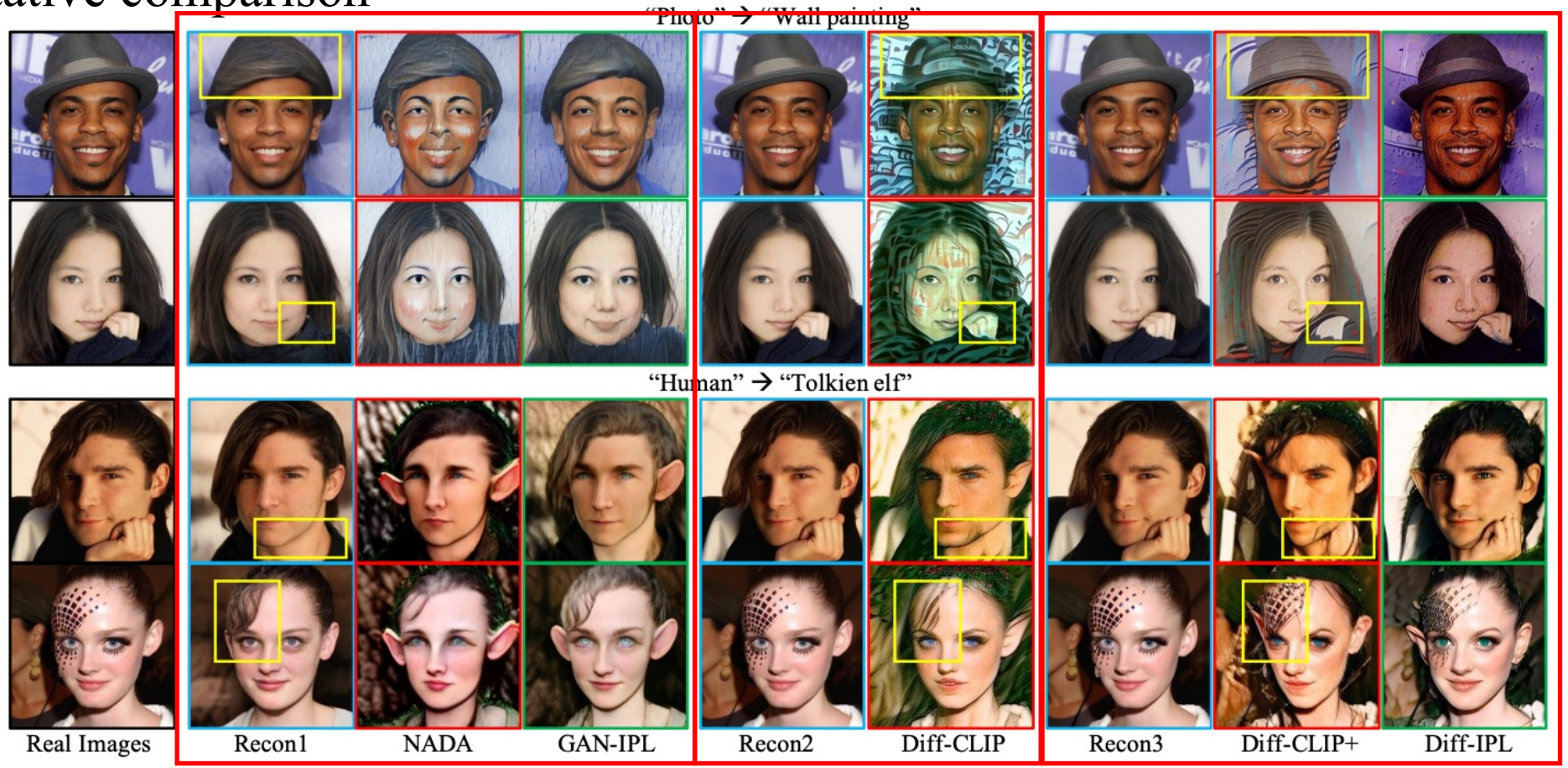
Quantitative comparison

Dataset	Source→Target	IS [41] (↑)		SCS [53] (↑)		ID [5, 12] (↑)		SIFID [42] (↓)						US (↑)
		NADA	IPL	NADA	IPL	NADA	IPL	NADA			IPL			
								R_1	R_2	R_3	R_1	R_2	R_3	
FFHQ [8]	Photo→Disney	2.721	**3.089**	0.407	**0.448**	0.782	**0.801**	2.776	3.136	3.670	**2.517**	**2.930**	**3.497**	82.6%
	Photo→Anime painting	2.450	**3.051**	0.324	**0.518**	0.666	**0.776**	2.956	1.811	1.242	**2.845**	**1.595**	**1.021**	79.3%
	Photo→Wall painting	2.183	**2.676**	0.439	**0.487**	0.594	**0.637**	1.944	1.220	1.331	**1.930**	**1.183**	**1.274**	80.9%
	Photo→Ukiyo-e	2.205	**2.974**	0.420	**0.506**	**0.775**	0.632	1.954	1.990	1.326	**1.165**	**1.255**	**0.878**	85.9%
	Human→Pixar character	2.703	**2.785**	0.379	**0.461**	0.757	**0.853**	0.793	0.932	**0.865**	**0.638**	**0.821**	1.092	86.7%
	Human→Tolkien elf	2.479	**2.778**	0.416	**0.491**	0.711	**0.772**	**0.632**	1.495	1.452	0.690	**0.637**	**0.701**	76.8%
	Human→Werewolf	2.619	**2.809**	0.399	**0.417**	0.642	**0.747**	1.969	1.846	1.967	**1.734**	**1.688**	**1.911**	72.7%
AFHQ [3]	Photo→Cartoon	6.505	**8.658**	0.407	**0.563**	0.925	**0.941**	2.708	2.672	3.870	**2.517**	**2.477**	**3.278**	87.6%
	Photo→Pointillism	5.419	**6.913**	0.224	**0.542**	0.775	**0.881**	7.081	5.288	7.142	**4.818**	**3.089**	**4.074**	78.5%
	Photo→Cubism	4.165	**6.450**	0.386	**0.463**	0.934	**0.943**	2.779	**2.938**	3.199	**2.431**	2.956	**2.284**	74.3%

IPL generates images with: (1) higher quality and diversity (IS), (2) better source-domain information preserving capability (e.g., structure (SCS) or identity (ID)), (3) more correct target-domain style (SIFID) and (4) more user study preference (US).

Experiments – *Real-world Image Translation*

Qualitative comparison

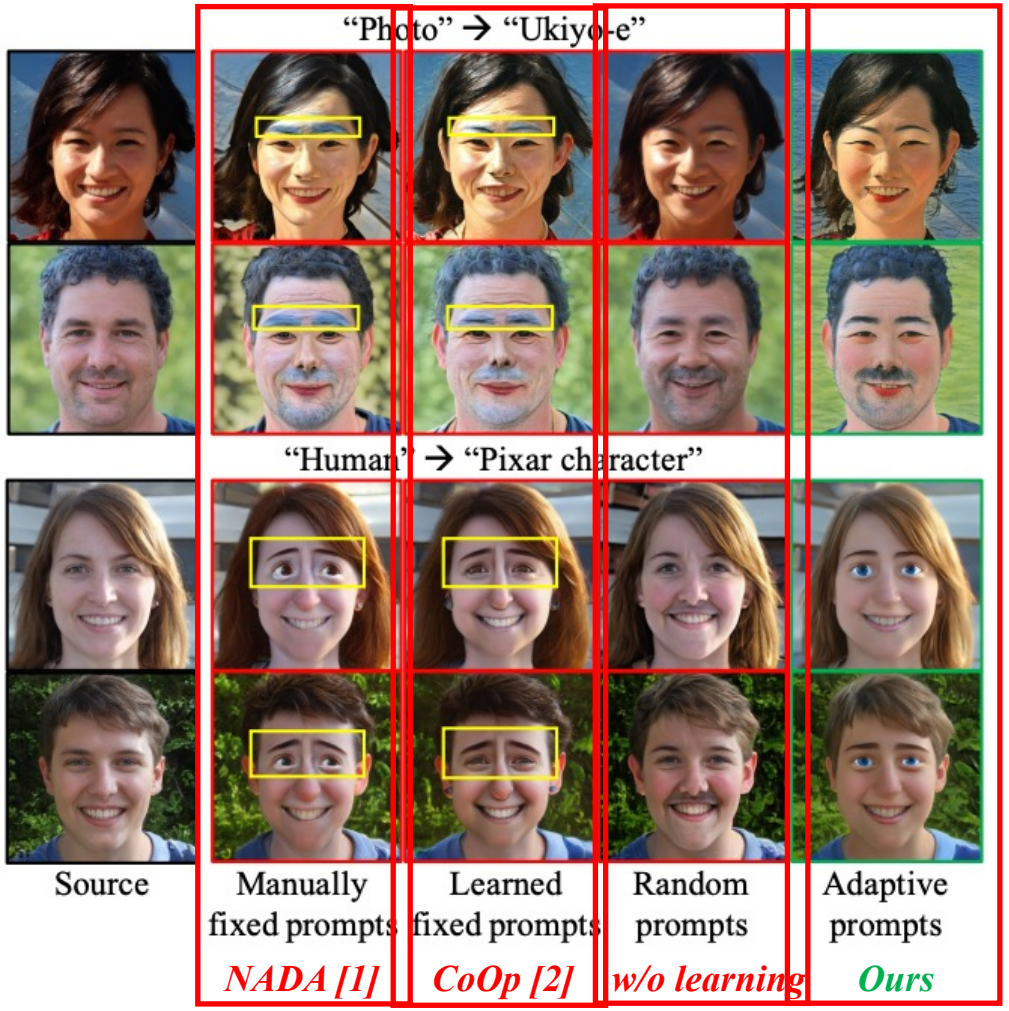


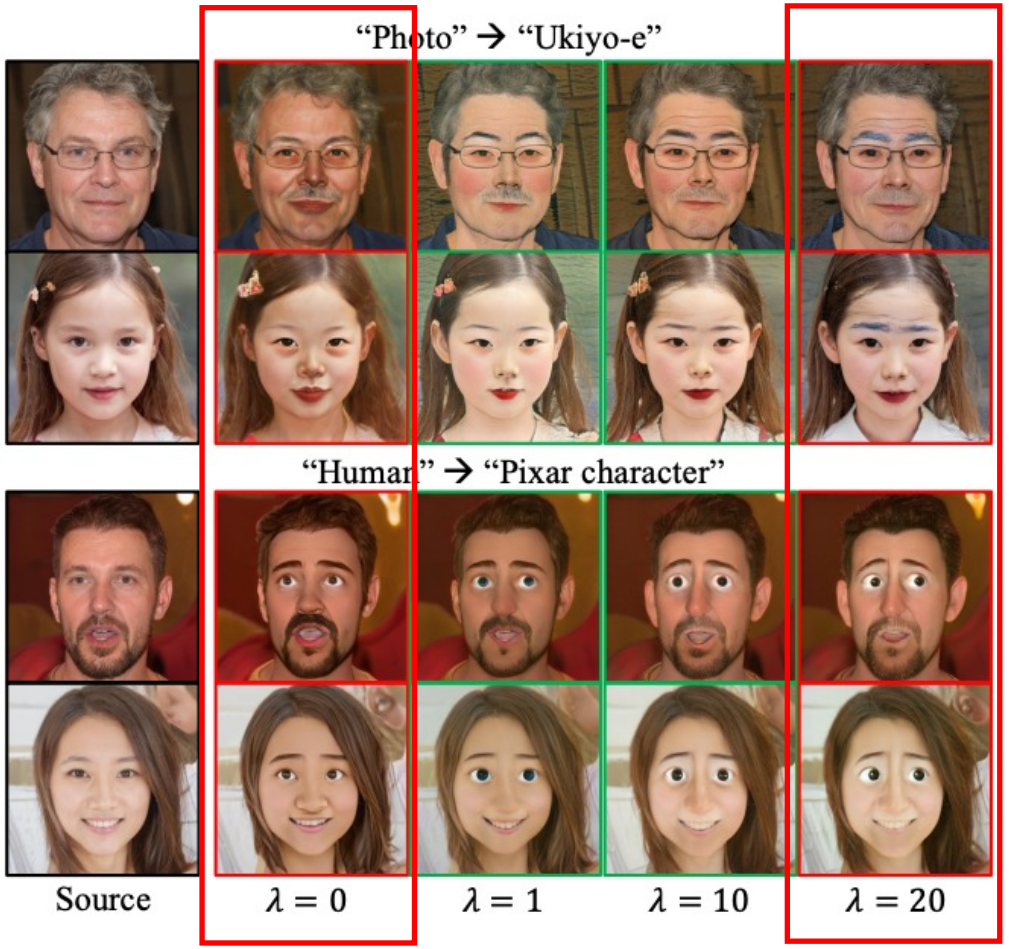
IPL is compatible with both GANs (GAN-IPL) and diffusion models (Diff-IPL). Diff-IPL inverses real-image more faithfully.

[1] Gal, Rinon, et al. "StyleGAN-NADA: CLIP-guided domain adaptation of image generators." *TOG* 2022.
[2] Kim, Gwanghyun, et al. "DiffusionCLIP: Text-guided diffusion models for robust image manipulation." *CVPR* 2022.
[3] Preechakul, Konpat, et al. "Diffusion Autoencoders: Toward a meaningful and decodable representation." *CVPR* 2022.

Experiments – *Ablation studies*

Prompt designing scheme & Loss term ratios





$$\mathcal{L} = \mathcal{L}_{\text{constr}} + \lambda \mathcal{L}_{\text{domain}}$$

Recommended range: [1, 10]

[1] Gal, Rinon, et al. "StyleGAN-NADA: CLIP-guided domain adaptation of image generators." *TOG* 2022.
[2] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *IJCV* 2022.

Thanks for watching!

For more details, please refer to our paper.

Paper

Code & Models