PuLID: Pure and Lightning ID Customization via Contrastive Alignment

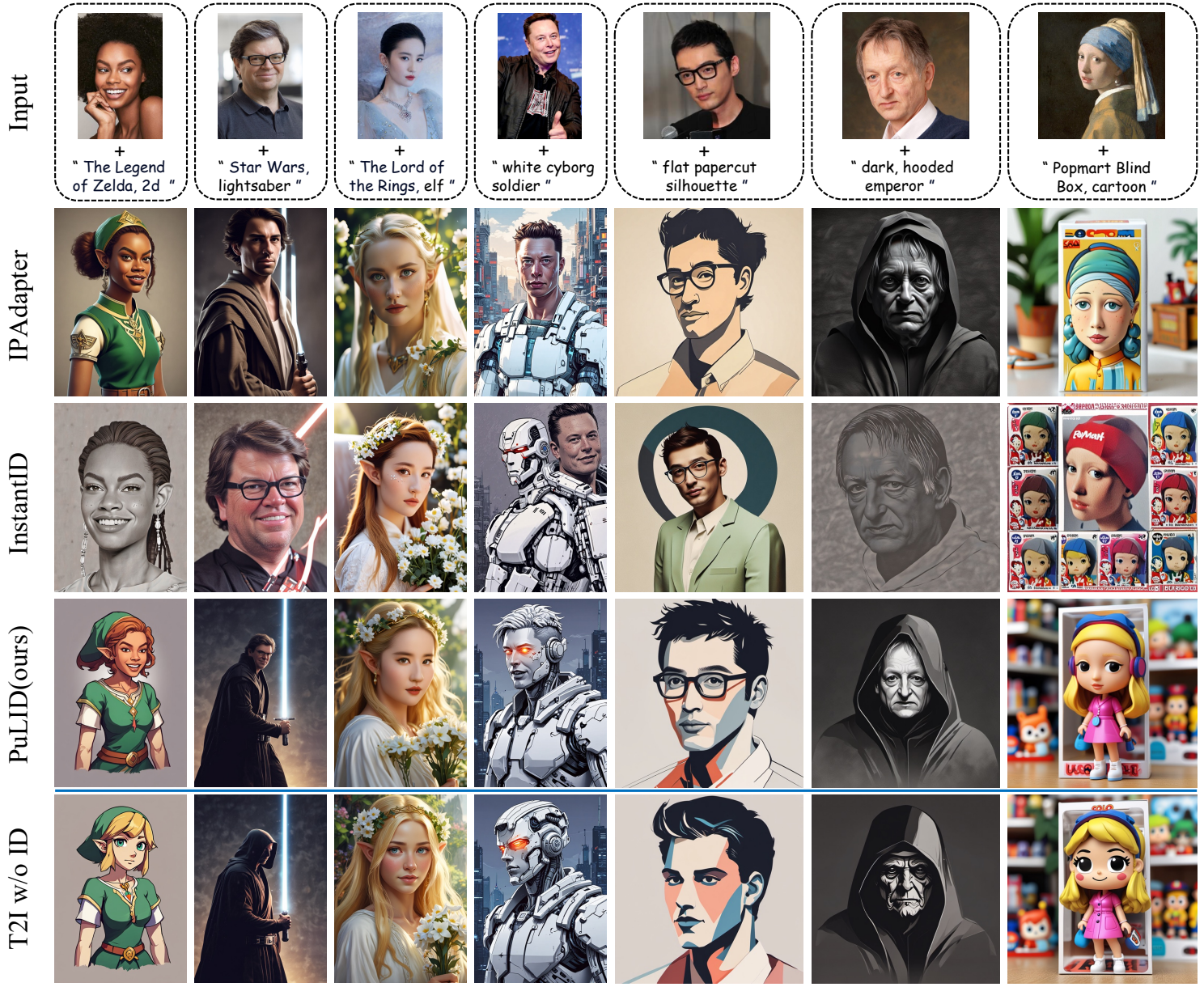
Zinan Guo[*1] Yanze Wu[*#] Zhuowei Chen Lang Chen Peng Zhang Qian He

ByteDance Inc

codes

* equal contribution, # corresponding author

Our Goal – Pure and Lightning ID Customization

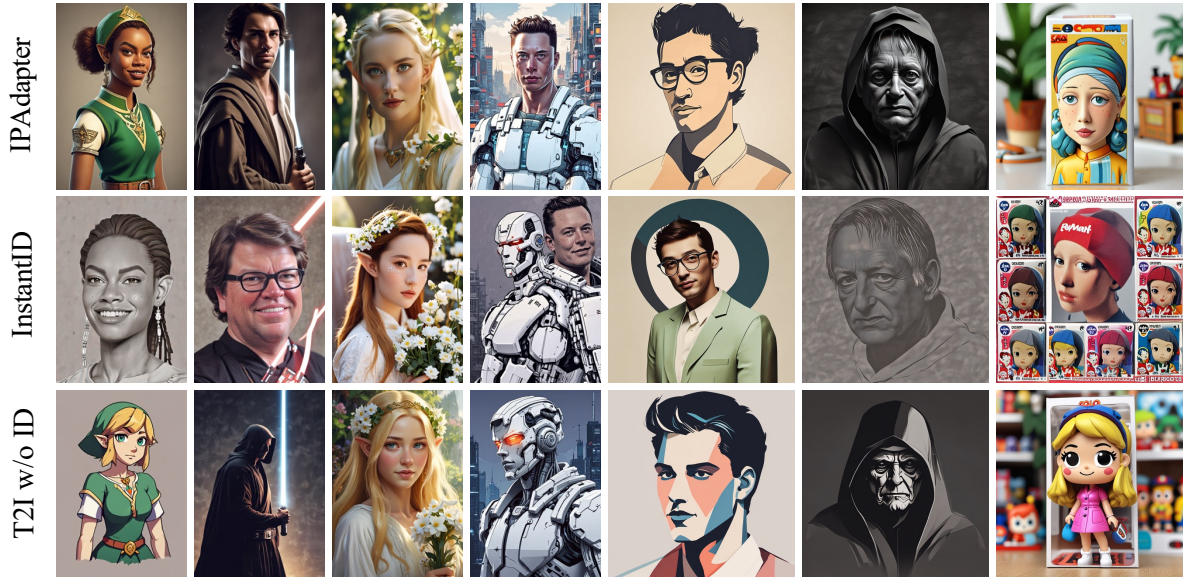


- **Lightning**: tuning-free, no need LoRA
- **Pure**: minimizing disruption to the original model after ID insertion
- **High ID fidelity**

Challenges

- Insertion of ID disrupts the original model's behavior

1. An ideal ID insertion should only alter ID-related aspects (e.g., face, hair), the ID-unrelated image emlements (e.g., b.g., lighting, composition, style) should be consistent with the original model



✗ Previous methods would alter ID-unrelated image elements when inserting ID

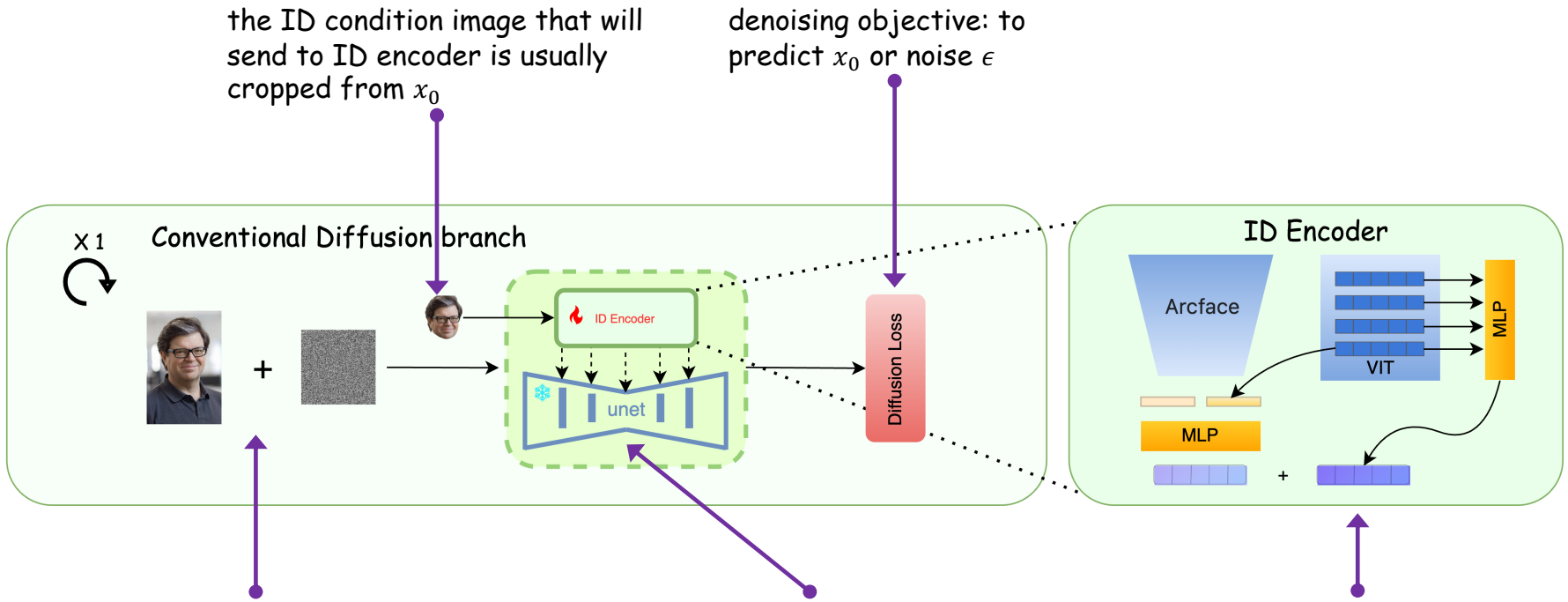
Models with higher ID fidelity tend to induce more severe disruption

2. After the ID insersion, it should still retain the ability of the original T2I model to follow prompts
   - which means altering ID attributes, orientation, and accessories via prompts
   - Previous effors includes
     - ➢ enhancing the encoder: IPAdapter and InstantID switch from CLIP-ViT to face rec. model to extract ID feature
     - ➢ constructing datasets grouped by ID to support non-reconstructive training: PhotoMaker

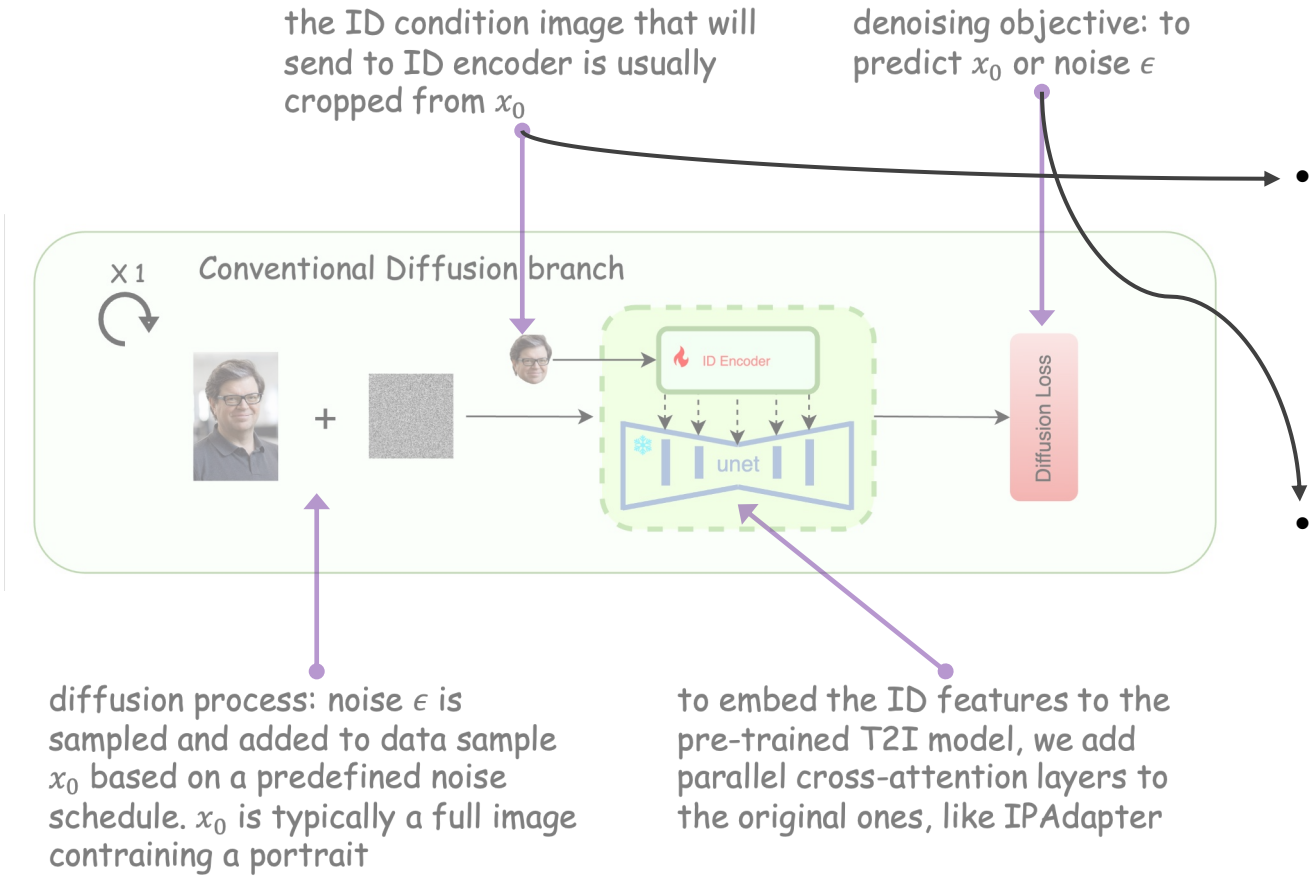
Challenges

- Lack of ID fidelity
  - Given our human sensitivity to faces, maintaining a high degree of ID fidelity is crucial for our task
  - Inspired by the GAN-based face generation methods, a straightforward idea for improving ID fidelity is to introduce ID loss within diffusion training
    - However, directly predict x_0 from x_t and calculate ID loss is inaccurate
      - the predicted x_0 is often noisy and flawed
      - the face recognition model is trained on photo-realistic images

Common Diffusion Training in ID Customization



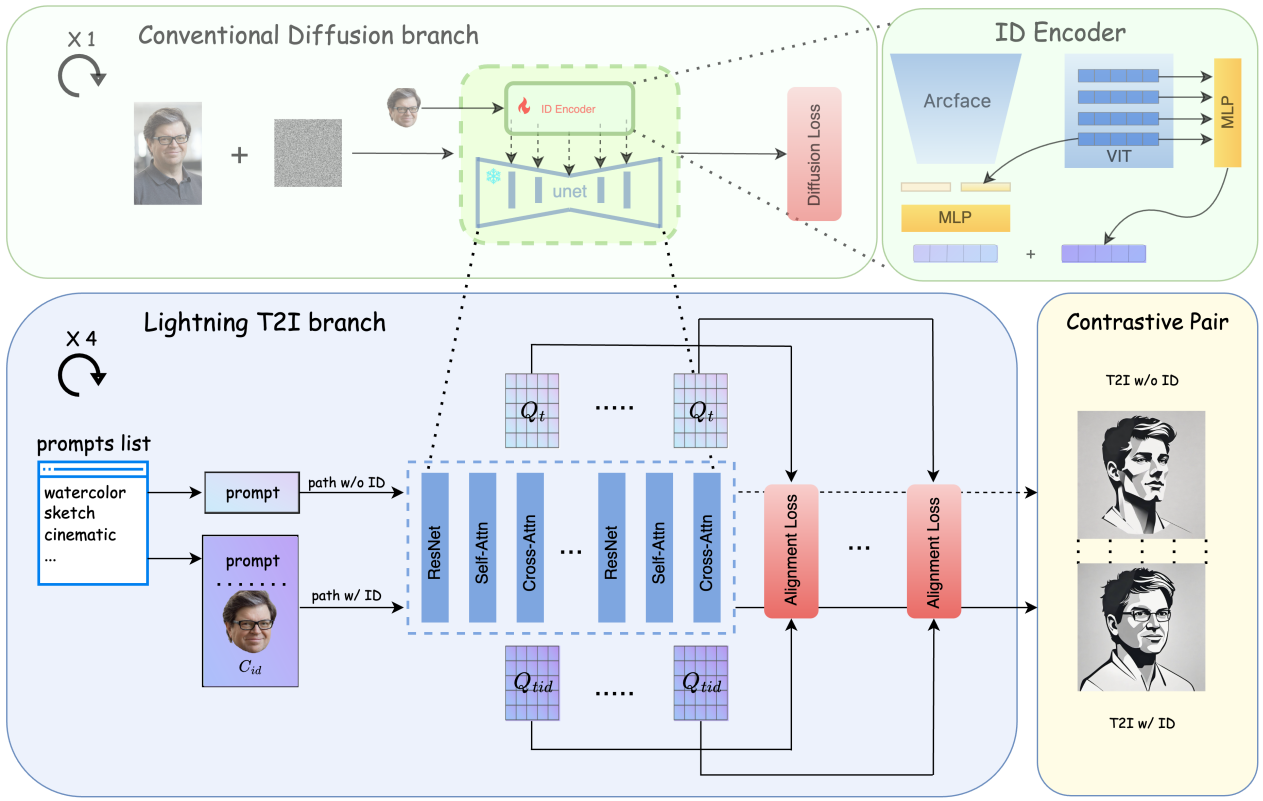
Common Training Paradigm is Flawed



- The ID condition aligns completely with the prompt and UNET features, implying it does not constitute contamination to the T2I model during training (but during testing, the prompt is usually conflict or misaligned with ID cond).

- To better reconstruct x_0, the model will use all the information from ID features (reason for copy-paste), as well as bias the training parameters towards the dataset distribution (reason for hard to changing style)

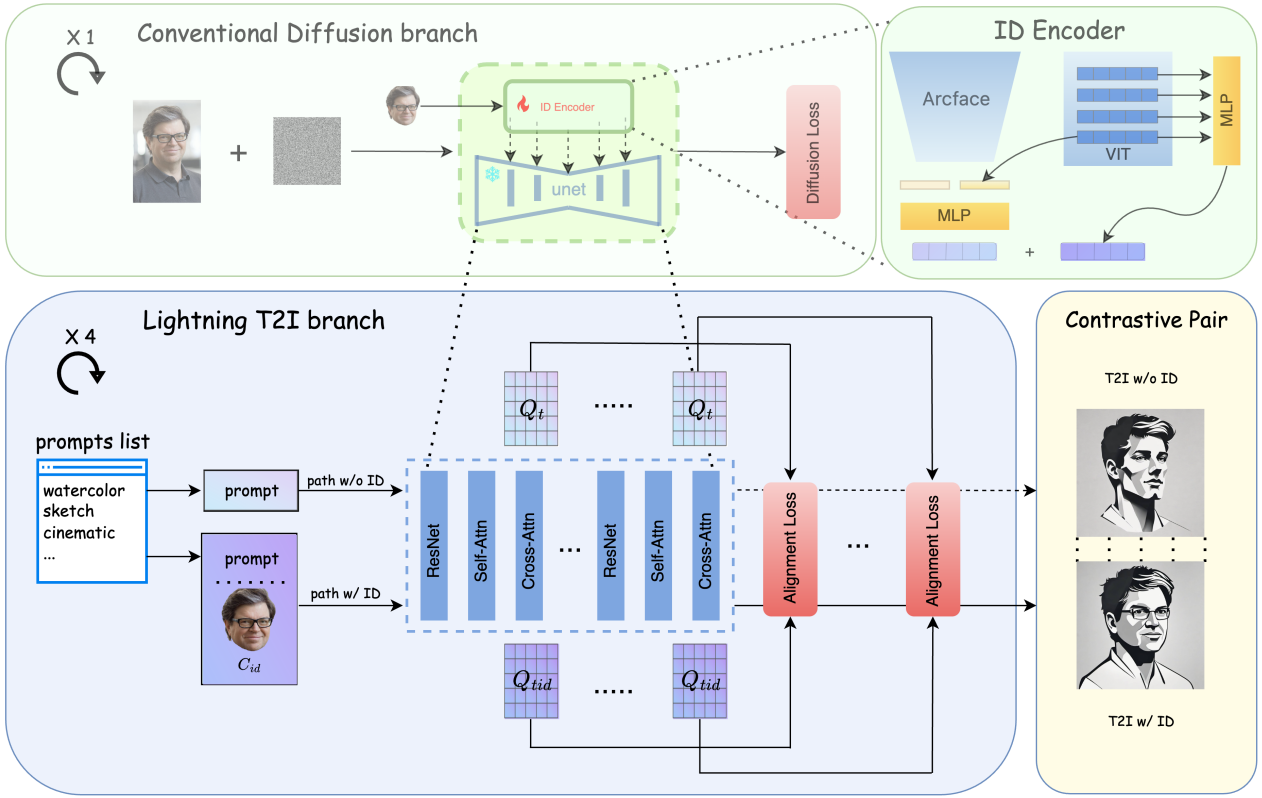
Pure ID Insertion via Contrastive Alignment



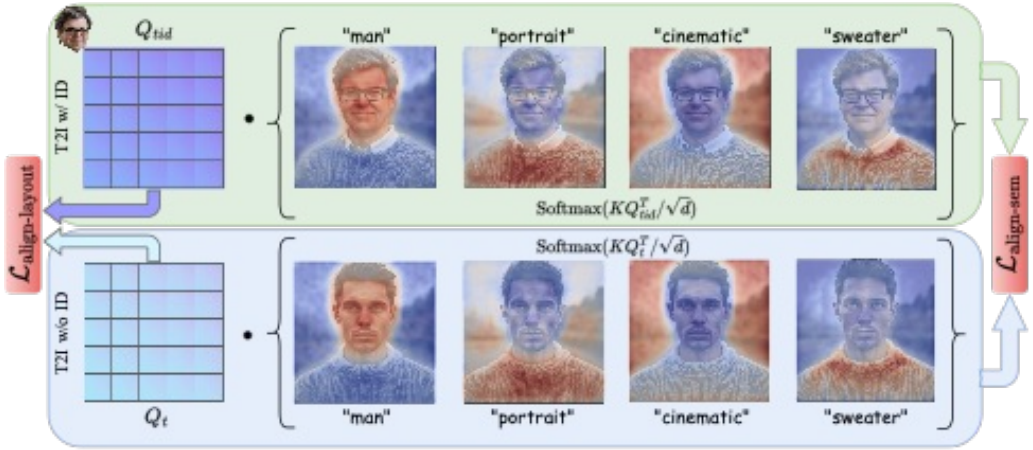
the Lighting T2I branch starts from pure noise and goes through the full iterative denoising steps until reaching x_0. Concretely, we employ SDXL-Lightning with 4 denoising steps

- To achieve uncontaminated ID insertion, we introduce a Lightning T2I training branch beyond the conventional diffusion-desnoising training branch
- We construct contrastive paths that start from the same prompt and initial latent. One path is conditioned only by the prompt, while the other path employs both the ID and the prompt as conditions
- By semantically aligning the UNET features on these two paths, the model will learn how to embed ID without impacting the behavior of the original model.

Pure ID Insertion via Contrastive Alignment (Cont.)



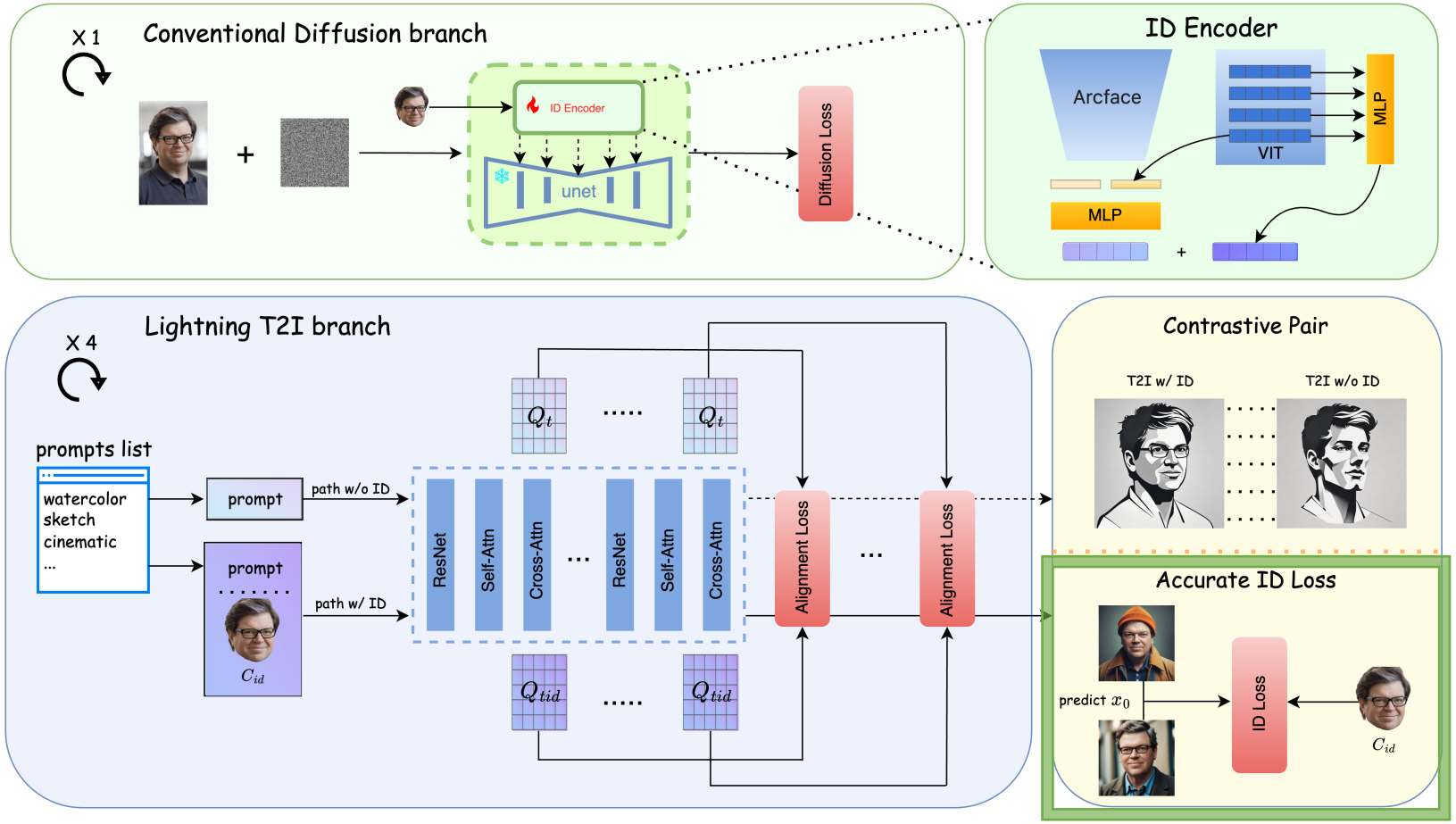
- To achieve uncontaminated ID insersion, we introduce a Lightning T2I training branch beyond the conventional diffusion-desnoising training branch
- We construct contrastive paths that start from the same prompt and initial latent. One path is conditioned only by the prompt, while the other path employs both the ID and the prompt as conditions
- By semantically aligning the UNET features on these two paths, the model will learn how to embed ID without impacting the behavior of the original model.



As illustrated in the left figure, our alignment loss consists of two components:

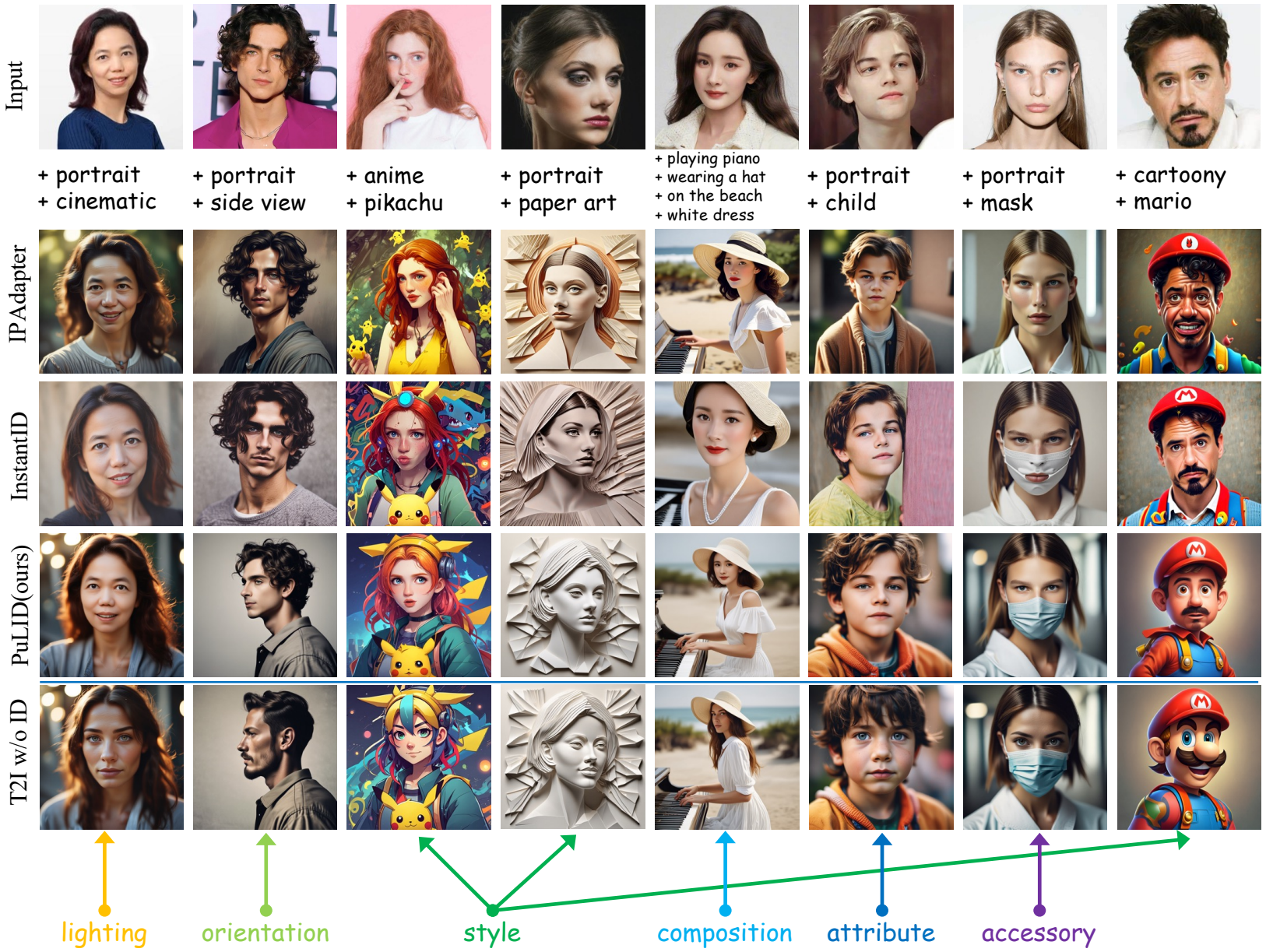
- semantic alignment loss $\mathcal{L}_{\text{align-sem}} = \left\| \text{Softmax}(\frac{KQ_{tid}^T}{\sqrt{d}})Q_{tid} - \text{Softmax}(\frac{KQ_t^T}{\sqrt{d}})Q_t \right\|_2$
- layout alignment loss $\mathcal{L}_{\text{align-layout}} = \|Q_{tid} - Q_t\|_2 .$

Optimizing ID Loss in a More Accurate Setting



- Thanks to the introduced Lightning T2I branch, we can swiftly generate an accurate x_0 conditioned on the ID from pure noise within 4 steps
- Calculation ID loss on this x_0, which is very close to the real-world data distribution, is evidently more precise

Qualitative Comparisons



Quantitative Comparisons

Table 1: **Quantitative comparisons.** *We observed that PhotoMaker shows limited compatibility with SDXL-Lightning, hence, we compare its performance on SDXL-base in this table.

	DivID-120			Unsplash-50		
	Face Sim.↑	CLIP-T↑	CLIP-I↑	Face Sim.↑	CLIP-T↑	CLIP-I↑
PhotoMaker*	0.271	26.06	0.649	0.193	27.38	0.692
IPAdapter	0.619	28.36	0.703	0.615	28.71	0.701
InstantID	0.725	28.72	0.680	0.614	30.55	0.736
PuLID (ours)	**0.733**	**31.31**	**0.812**	**0.659**	**32.16**	**0.840**

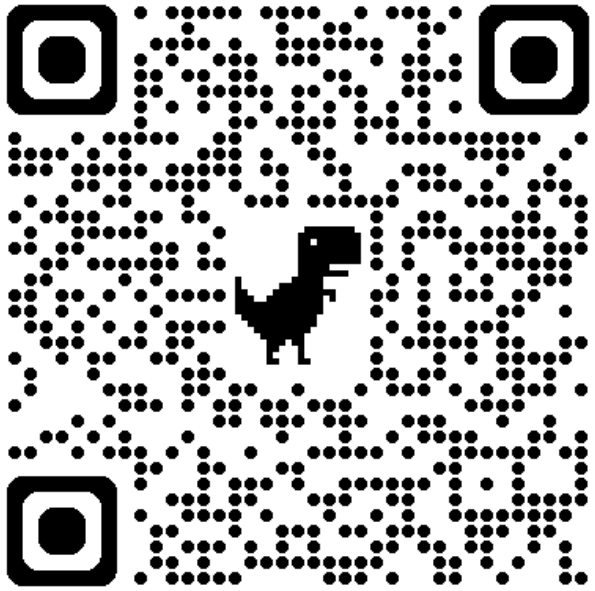
CLIP-I: quantify the CLIP image similarity between two images before and after the ID insertion. A higher CLIP-I metric indicates a smaller modification in image after ID insertion, suggesting a lower degree of disruption to the original model's behavior.

Thanks for Watching ☺

ByteDance

NEURAL INFORMATION PROCESSING SYSTEMS

PuLID: Pure and Lightning ID Customization via Contrastive Alignment



Codes, HuggingFace demos