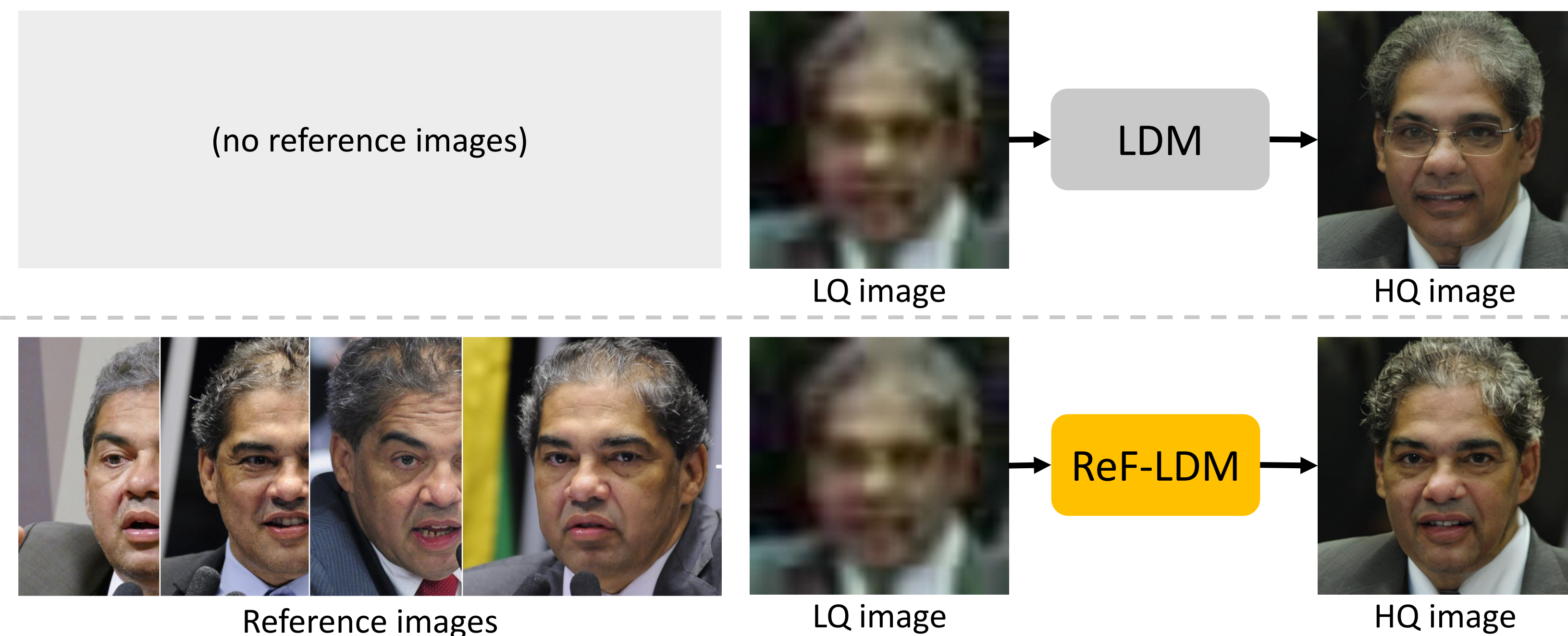# ReF-LDM: A Latent Diffusion Model for Reference-based Face Image Restoration

Chi-Wei Hsiao, Yu-Lun Liu, Cheng-Kun Yang,
Sheng-Po Kuo, Yucheun Kevin Jou, Chia-Ping Chen

MEDIATEK nycu NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Project Page

Motivation

(no reference images)

LQ image → LDM → HQ image

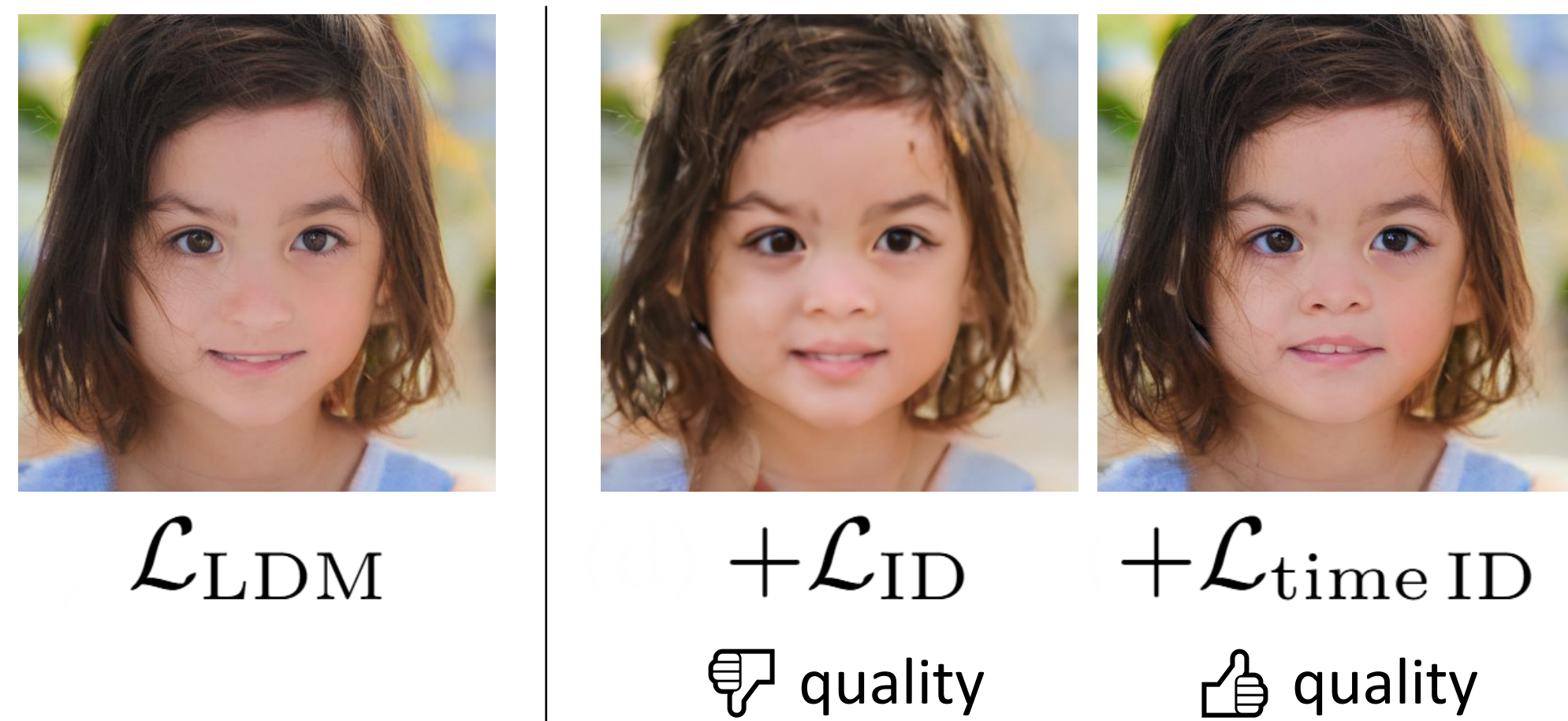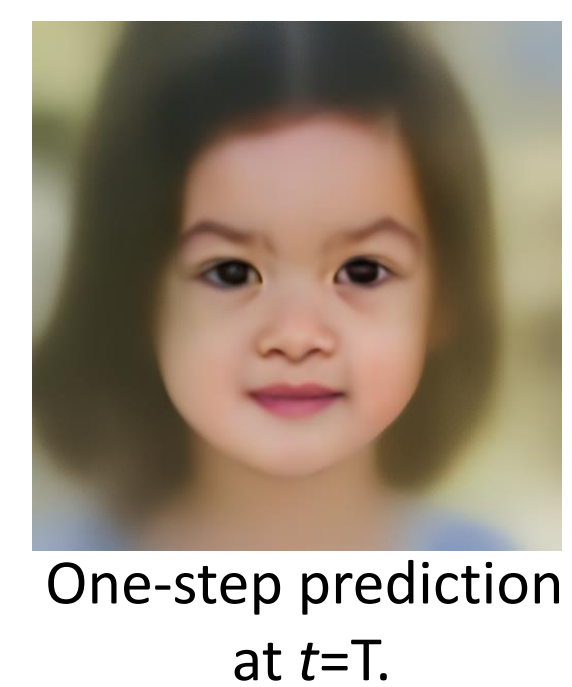Reference images | LQ image → ReF-LDM → HQ image

Challenge: Reference images are spatially unaligned

ReF-LDM
1. CacheKV: cache and reuse KV of reference images
2. Timestep-scaled Identity Loss
3. FFHQ-Ref dataset

Timestep-scaled Identity Loss

- Identity loss ($\mathcal{L}_{ID}$) worsens the image quality of a diffusion model.
- Possible reason:

  $\mathcal{L}_{ID}$ is feature distance of a pre-trained ArcFace model.

  Diffusion training uses one-step prediction.

  One-step prediction at larger timestep is OOD for ArcFace.

- Solution: $\mathcal{L}_{\text{time ID}} = \sqrt{\bar{\alpha}_t} \cdot \mathcal{L}_{ID}$ (downscale $\mathcal{L}_{ID}$ for a larger t)

One-step prediction at $t=T$.

$\mathcal{L}_{LDM}$ | $+\mathcal{L}_{ID}$ | $+\mathcal{L}_{\text{time ID}}$
(thumbs down) quality | (thumbs up) quality

Loss	IDS↑	NIQE↓
$\mathcal{L}_{LDM}$	0.52	4.56
$\mathcal{L}_{LDM} + \mathcal{L}_{ID}$	0.69	6.56
$\mathcal{L}_{LDM} + \mathcal{L}_{\text{time ID}}$	0.65	4.38

Ablation.

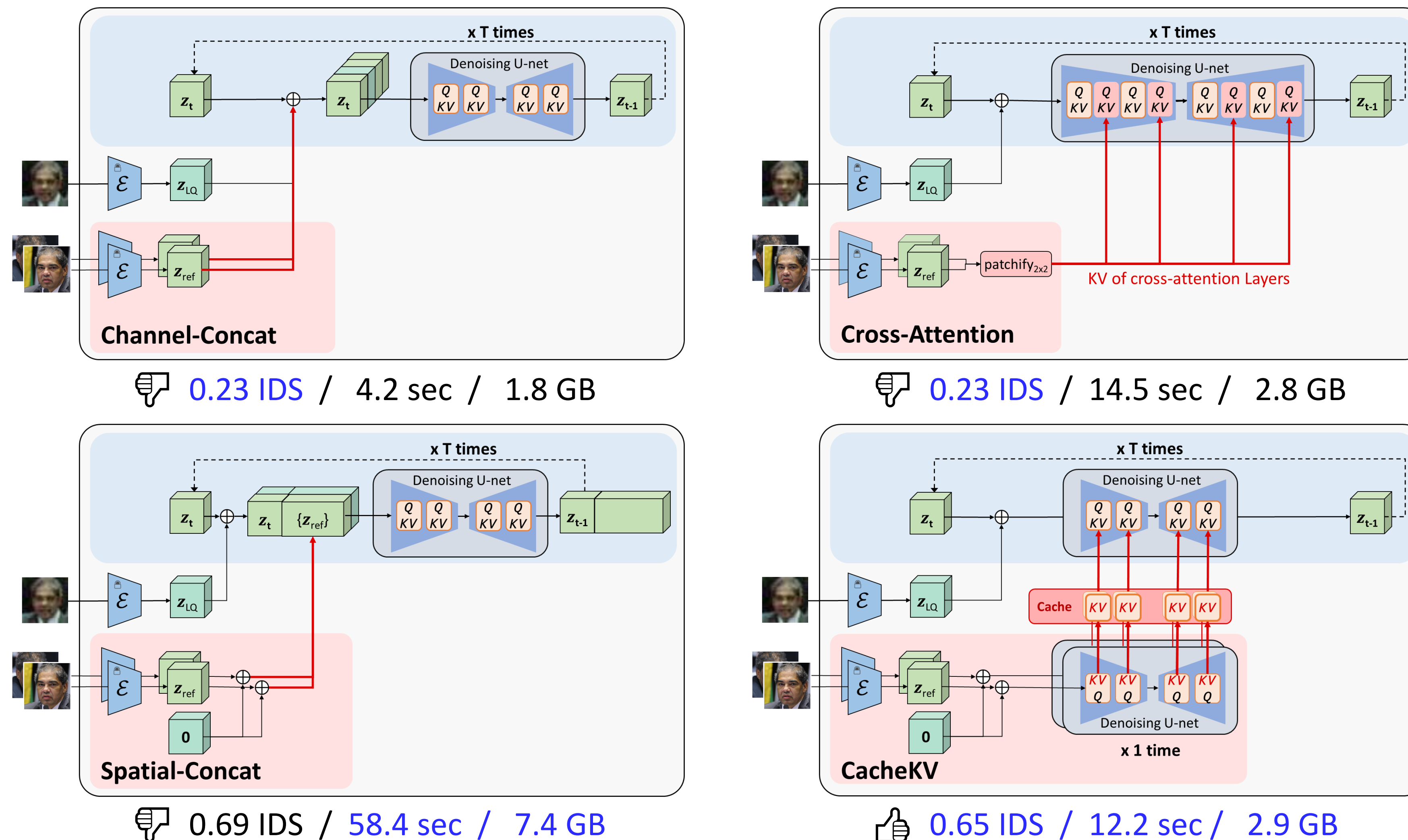| Scale for ID loss	IDS↑	NIQE↓
$\sqrt{\bar{\alpha}_t}$	0.65	4.38
$\mathbf{1}_{t<100}$	0.52	4.55
$\mathbf{1}_{t<500}$	0.61	4.44

Scaling factor design choices.

Model Pipeline

- Adapt LDM to condition on **LQ image** and **reference images**.
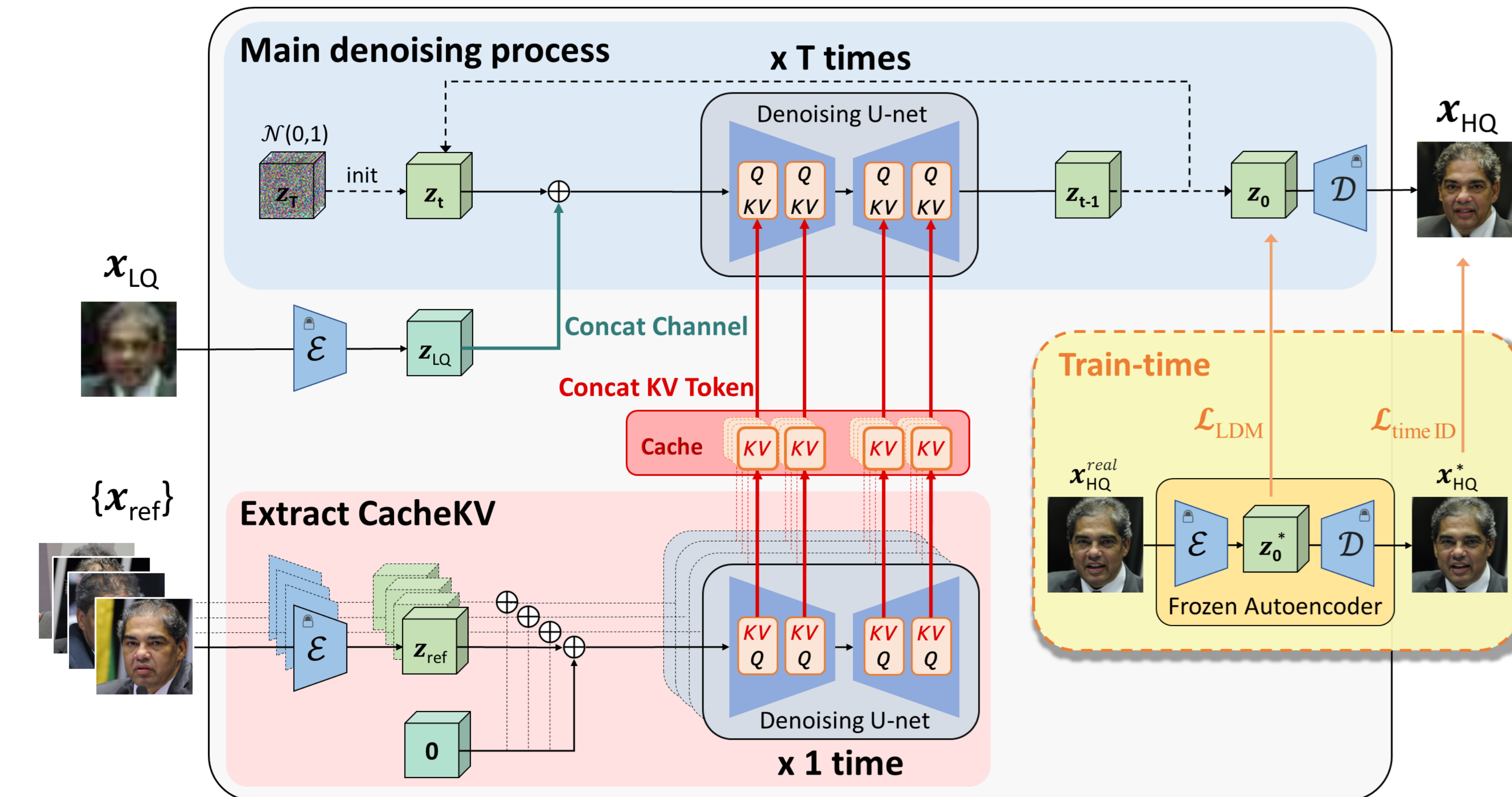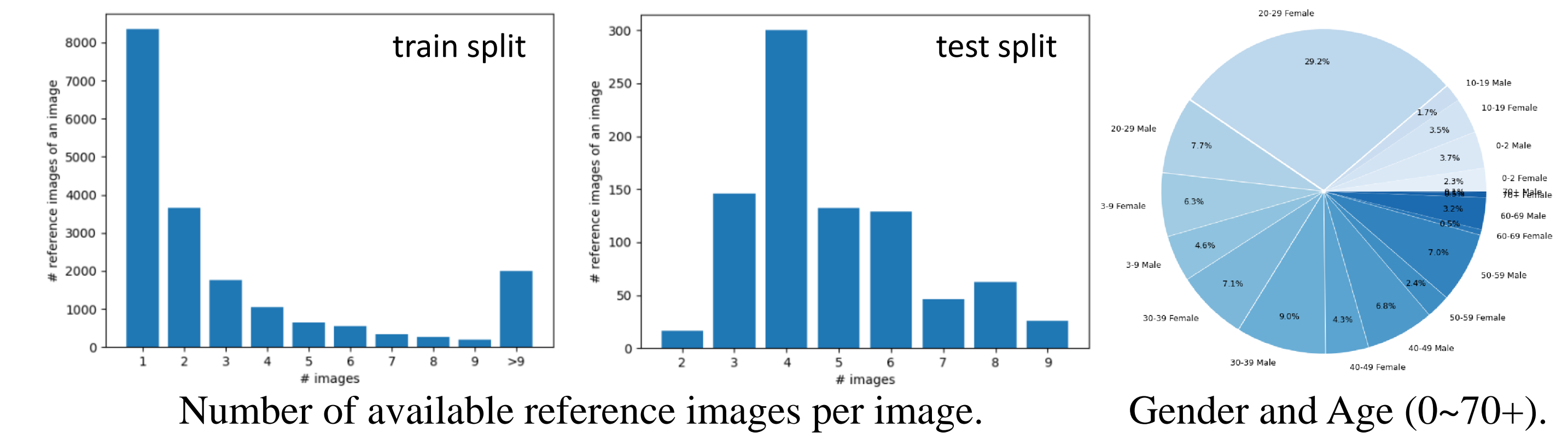

CacheKV: Make LDM Condition on Reference Images

Ablation (*Low IDS = failure to use references. *Tested with five reference images on a single GTX 1080.)
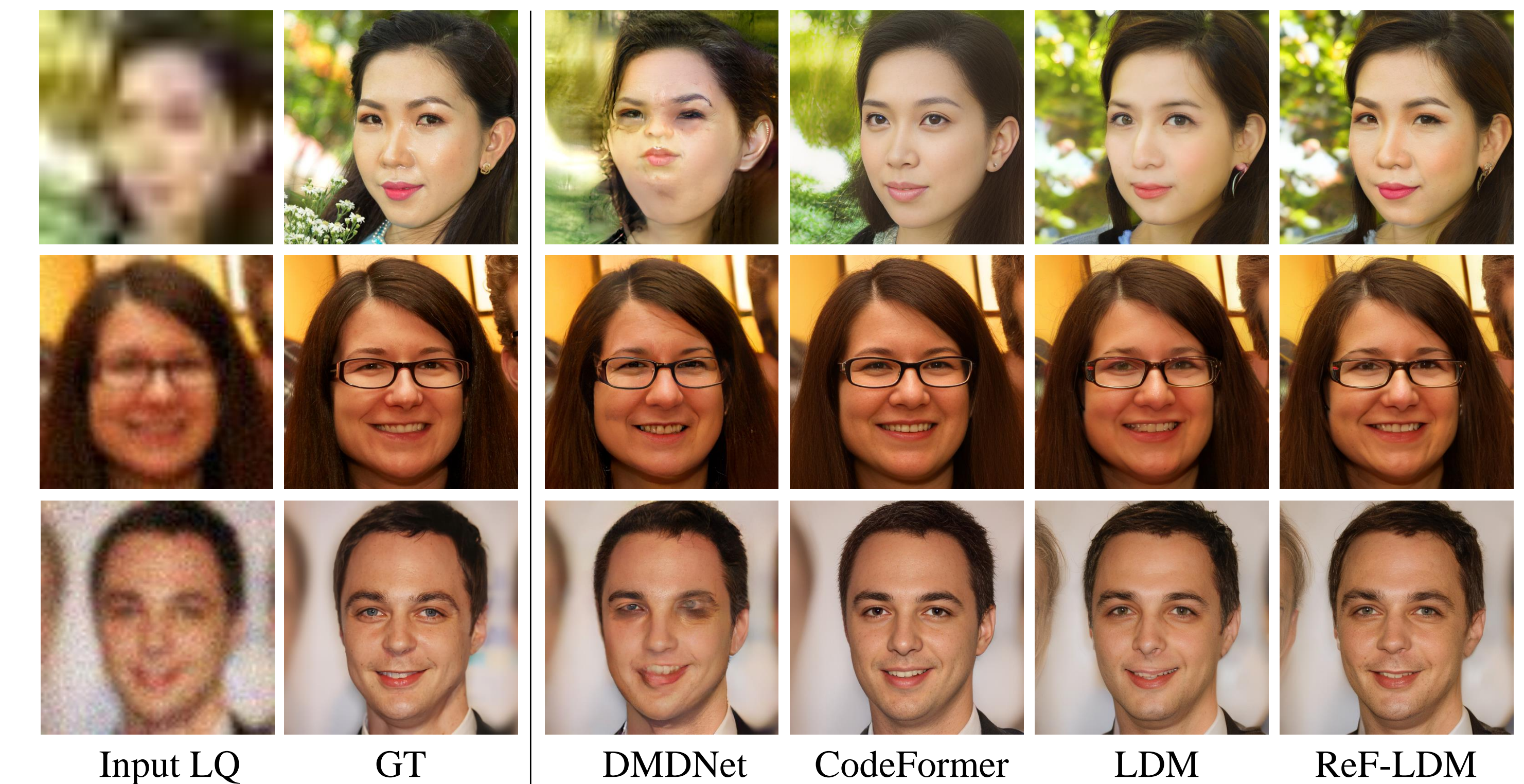

Channel-Concat
(thumbs down) 0.23 IDS / 4.2 sec / 1.8 GB

Cross-Attention
(thumbs down) 0.23 IDS / 14.5 sec / 2.8 GB

Spatial-Concat
(thumbs down) 0.69 IDS / 58.4 sec / 7.4 GB

CacheKV
(thumbs up) 0.65 IDS / 12.2 sec / 2.9 GB

FFHQ-Ref Dataset

- High-quality face images with corresponding reference images.

Dataset	With reference	Licensed	Quality	Images	Identities
FFHQ [12]		✓	✓	70,000	-
CelebRef-HQ [16]	✓	✓	✓	10,555	1,005
FFHQ-Ref	✓	✓	✓	**20,405**	**6,523**

Number of available reference images per image. | Gender and Age (0~70+).

Results

Input LQ | GT | DMDNet | CodeFormer | LDM | ReF-LDM

	FFHQ-Ref-Severe				FFHQ-Ref-Moderate				CelebA-Test-Ref		
	IDS↑	fLPIPS↓	LPIPS↓	FID↓	IDS↑	fLPIPS↓	LPIPS↓	FID↓	IDS↑	fLPIPS↓	LPIPS↓
CodeFormer [32]	0.323	0.108	0.398	51.51	0.760	0.084	0.301	38.78	0.660	0.092	0.340
VQFR [5]	0.308	0.112	0.415	52.96	0.659	0.089	0.324	36.77	0.558	0.096	0.352
DAEFR [26]	0.294	0.118	0.435	49.08	0.614	0.093	0.333	33.86	0.491	0.101	0.367
LDM	0.231	0.125	0.453	34.40	0.753	0.095	0.344	32.16	0.663	0.093	0.368
DMDNet [16]†	0.185	0.162	0.511	72.66	0.810	0.096	0.348	36.60	0.752	0.097	0.362
ReF-LDM	0.676	0.110	0.429	37.60	0.840	0.088	0.332	33.05	0.779	0.093	0.368

†As DMDNet encounters landmark detection failures and fails to yield results for 214/857, 29/857, and 488/2,533 images on the three benchmarks respectively, we compute the metrics for DMDNet using the remaining images.