Supplementary Material for IConFace: Identity-Structure Asymmetric Conditioning for Unified Reference-Aware Face Restoration

Axi Niu*, **Jinyang Zhang***, **Senyan Qing**
Northwestern Polytechnical University
nax@nwpu.edu.cn zhangjinyang@mail.nwpu.edu.cn qingsenyan@nwpu.edu.cn
Homepage: https://cosmicrealm.github.io/IConFace/

Supplementary Overview

This supplement provides protocol details, implementation specifications, and diagnostic evidence supporting the main paper: (i) dataset, training, inference, comparison-method, and metric protocols; (ii) reference–GT gap statistics, target-state distortion/structure checks, and GT-AdaFace analysis; and (iii) extended reference-aware, no-reference, and ablation visual comparisons. The main paper reports the compact method and benchmark summary, while this supplement documents the implementation choices and additional quantitative/qualitative evidence behind those results.

Protocol and Implementation Details

Training data. IConFace is trained on FFHQ-Ref, a reference-aware extension of FFHQ. FFHQ contains 70,000 aligned 1024×1024 face images with broad variation in age, ethnicity, background, accessories, and image conditions. FFHQ-Ref groups FFHQ images by ArcFace-predicted identity and provides reference mappings for images that have same-identity partners. The released FFHQ-Ref reference graph contains 20,405 usable high-quality images; our training split is the trainval mapping, which contains 6,373 identity groups and 19,548 usable images with at least one same-identity reference. During training, the target image is degraded online with mixed blind degradations, while same-identity images from the mapping serve as optional protocol references.

Evaluation data. Reference-aware evaluation uses CelebA-Test-Ref, FFHQ-Ref Moderate, FFHQ-Ref Severe, and CelebHQRef100. CelebA-Test-Ref contains 2,533 test samples with paired LQ/GT images and same-identity references from CelebA-HQ/CelebAMask-HQ. FFHQ-Ref Moderate and FFHQ-Ref Severe each contain 857 FFHQ-Ref test targets with fixed same-identity references and two synthetic degradation levels. CelebHQRef100 is a compact diagnostic benchmark built from 100 high-quality identity groups. The identities are selected deterministically from the sorted source identity folders, the first sorted image is used as target/GT, up to three remaining same-identity images are used as protocol references, and the target is degraded with mixed BSRGAN and Real-ESRGAN style degradations using downsampling scales from 4 to 10. The degradation seed is fixed as $42 + i$ for identity index i to make the generated LQ images deterministic. CelebHQRef100 is not used for training, checkpoint selection, or prompt tuning. We verified that no released CelebHQRef100 image path or protocol identity folder appears in the FFHQ-Ref trainval mapping; the split is used only as a compact diagnostic benchmark. No-reference evaluation uses CelebA-Test (3,000 paired synthetic samples), LFW (1,711), CelebChild (360), WebPhoto (407), and Wider-Test (970); LFW, CelebChild, WebPhoto, and Wider-Test are treated as real-world sets without paired ground truth.

Backbone, training, and inference. IConFace uses the FLUX.2-klein-base-4B restoration backbone and adds rank 16 LoRA adapters together with the identity and degraded-structure modules described in the main paper. The final optimization setting uses 512×512 crops, fixed learning rate `1e-5`, batch size 1, and gradient accumulation 4. Degradation strength is sampled from 0–16 with bucket probabilities 0.5, 0.3, and 0.2 for ranges 0–3, 4–8, and 9–16. All reported results use 12 sampling steps, guidance scale 4.0, base seed 42, and 512 resolution.

Reference sampling and conditioning. Training mixes reference availability so the same checkpoint can operate with or without references: 30% of samples use no reference, 30% use one, 20% use two, and 20% use three. When fewer than three references are available, we use the actual number without duplication. The global identity pathway uses AdaFace IR50 embeddings with norm-only multi-reference aggregation and temperature 1.0. The degraded structure pathway uses degraded strength 1.0 and low-rank degraded cross-attention rank 16. No-reference inference removes reference tokens and uses the degraded-image AdaFace fallback only as weak forward conditioning.

The implemented objective uses the same losses as the main paper with explicit scalar multipliers. Eq. (8) uses uniform flow-matching timestep weight, $w(\sigma) = 1$ (`loss_weighting=none`), and scalar multiplier $\alpha_{\mathrm{fm}} = 0.75$. The AdaFace identity multiplier is $\lambda_{\mathrm{id}} = 0.30$, the hard target stabilizer uses $\lambda_h = 0.25$, and the sigma floor is $\omega_{\min} = 0.25$:

$$\mathcal{L} = \alpha_{\mathrm{fm}}\mathcal{L}_{\mathrm{fm}} + \lambda_{\mathrm{id}}\omega(\sigma)[(1-\lambda_h^\star)\mathcal{L}_{\text{ref-id}} + \lambda_h^\star\mathcal{L}_{\mathrm{hard}}], \quad (1)$$

*These authors contributed equally.

where $\lambda_h^\star = \lambda_h(1 - \cos(e_{\text{ref}}, e_{\text{gt}}))$ and $\omega(\sigma) = \max(1 - \sigma, \omega_{\min})^2$. Legacy sigma-threshold launcher fields are ignored by the final identity loss; the only active sigma control is the floor $\omega_{\min} = 0.25$ above.

Token packing and RoPE. All image and text conditions use the four-axis FLUX.2 rotary position layout (t, h, w, l). Scene latents are packed from (B, C, H, W) to (B, HW, C) with ids $(0, h, w, 0)$, while text tokens use ids $(0, 0, 0, l)$. The degraded image condition is assigned a separate temporal group with $t = 2$. Reference images are packed as grouped visual tokens with $t = 10 + r$ for reference index r, which lets the transformer distinguish the generated scene, degraded observation, and each reference without changing the spatial (h, w) axes. For a token i with position id $p_i = (t_i, h_i, w_i, l_i)$, the query and key vectors are split over axes $a \in \{t, h, w, l\}$ with dimensions $d_a = [32, 32, 32, 32]$. For each two-channel pair m on axis a, the RoPE computation is

$$\omega_{a,m} = \theta^{-2m/d_a}, \qquad \theta = 2000, \tag{2a}$$

$$R(\rho, \omega) = \begin{bmatrix} \cos(\rho\omega) & -\sin(\rho\omega) \\ \sin(\rho\omega) & \cos(\rho\omega) \end{bmatrix}, \tag{2b}$$

$$\tilde{q}_{i,a,m} = R(p_i^a, \omega_{a,m}) q_{i,a,m}, \tag{2c}$$

$$\tilde{k}_{j,a,m} = R(p_j^a, \omega_{a,m}) k_{j,a,m}, \tag{2d}$$

$$\text{Attn}_{ij} \propto \exp\left(\frac{\tilde{q}_i^\top \tilde{k}_j}{\sqrt{d}}\right). \tag{2e}$$

This construction preserves the ordinary spatial axes while using the temporal axis to mark scene, degraded, and reference token groups.

Comparison methods. Reference-aware baselines are DMDNet, ReF-LDM, RestorerID, InstantRestore, and FaceMe; they are evaluated with the same protocol references whenever supported. Blind baselines are GFP-GAN, VQFR, CodeFormer, RestoreFormer++, and DAEFR, evaluated in the empty-reference setting. DMDNet learns dual memory dictionaries; ReF-LDM uses latent diffusion with high-quality references; RestorerID injects single-reference identity through an ID-preserving adapter; InstantRestore uses shared-image attention for single-step personalized restoration; and FaceMe extracts identity prompts from references. For blind baselines, GFP-GAN uses a GAN facial prior, VQFR uses a vector-quantized dictionary, CodeFormer predicts codebook entries with a transformer, RestoreFormer++ uses reconstruction-oriented priors, and DAEFR couples LQ evidence with high-quality codebook priors. References for these methods are provided in the main paper.

Metrics. For each reference-aware sample, let $\hat{I}_i$ be the restored image and R_i^1 be the first protocol reference. Ref-ArcFace and Ref-AdaFace are dataset averages of cosine similarities

$$\text{RefMetric} = \frac{1}{N} \sum_{i=1}^{N} \frac{E(\hat{I}_i)^\top E(R_i^1)}{\|E(\hat{I}_i)\|_2 \, \|E(R_i^1)\|_2}, \tag{3}$$

where E is the corresponding frozen ArcFace or AdaFace encoder. We also report GT-AdaFace in the analysis section as the cosine similarity between $\hat{I}_i$ and the paired target image G_i. PSNR, SSIM, and LPIPS are computed only on paired synthetic benchmarks and are treated as target-state distortion/structure checks rather than primary real-world quality metrics. MUSIQ, CLIP-IQA, and MANIQA are learned no-reference perceptual quality metrics; each restored image is scored independently and the dataset mean is reported. We use these metrics together so that reference-aligned identity consistency, target-state distortion, and perceptual quality can be read separately.

Table 1: Paired no-reference CelebA-Test distortion metrics. Main-paper no-reference results emphasize learned perceptual metrics; this table reports the standard paired distortion metrics for completeness.

Method	PSNR	SSIM	LPIPS
CodeFormer	25.146	0.685	0.227
GFP-GAN	25.105	0.696	0.241
VQFR	23.233	0.658	0.244
RestoreFormer++	25.313	0.685	0.226
DAEFR	22.591	0.628	0.250
IConFace	22.291	0.635	0.288

Reproducibility details. For reproducibility, we keep the restoration prompt, random seed, sampling steps, guidance scale, dataset split files, metric scripts, and checkpoint identifiers fixed across all reported evaluations. The released code package is organized around the inference and evaluation paths used for the main-paper tables and qualitative panels.

Reference-Aligned Identity Metric Analysis

Reference-aware restoration has two identity-related observations: the degraded target image, whose clean counterpart defines the target state, and one or more same-identity references, which provide external identity evidence. These observations are not spatially aligned and may differ in pose, expression, illumination, age, makeup, occlusion, or local facial state. Therefore, a GT-aligned identity score and a reference-aligned identity score do not always measure the same property. GT-aligned identity measures resemblance to the target-state image, while Ref-ArcFace and Ref-AdaFace measure consistency with the fixed reference evidence supplied by the protocol.

This distinction is important for interpreting reference-aware results. If the reference and GT are nearly identical, GT-based and reference-based identity scores tend to agree. When the reference and GT are only moderately aligned, however, a GT-only score can penalize a restoration that correctly uses reference identity evidence but does not reproduce the exact target-state appearance. We therefore treat Ref-AdaFace as the primary reference-aware identity metric, report perceptual quality metrics alongside it, and use qualitative comparisons to verify that the restored face still follows the degraded-image structure rather than copying the reference pose or expression.

Table 2: Identity similarity between the first protocol reference and the GT target over the full reference-aware evaluation splits. The last three columns report the percentage of samples whose Ref_1–GT AdaFace similarity falls below each threshold.

Dataset	N	Ref_1–GT ArcFace	Ref_1–GT AdaFace	AdaFace < 0.5	AdaFace < 0.6	AdaFace < 0.7
CelebA-Test-Ref	2533	0.616 ± 0.093	0.605 ± 0.104	15.52%	49.11%	80.77%
FFHQ-Ref Moderate	857	0.687 ± 0.084	0.663 ± 0.099	4.55%	23.45%	63.36%
FFHQ-Ref Severe	857	0.687 ± 0.084	0.663 ± 0.099	4.55%	23.45%	63.36%
CelebHQRef100	100	0.653 ± 0.107	0.640 ± 0.113	16.00%	34.00%	73.00%

Table 3: Paired target-state checks on the reference-aware benchmarks. SSIM is the direct structure check; PSNR and LPIPS give distortion context.

Dataset	Method	PSNR	SSIM	LPIPS
CelebA-Test-Ref	DMDNet	25.535	0.696	0.253
	ReF-LDM	23.901	0.638	0.268
	RestorerID	24.744	0.662	0.279
	InstantRestore	25.400	0.702	0.219
	FaceMe	25.947	0.698	0.253
	IConFace	22.327	0.632	0.277
FFHQ-Ref Moderate	DMDNet	25.844	0.726	0.234
	ReF-LDM	23.971	0.664	0.232
	RestorerID	24.924	0.691	0.240
	InstantRestore	25.452	0.727	0.213
	FaceMe	26.335	0.733	0.173
	IConFace	22.755	0.671	0.218
FFHQ-Ref Severe	DMDNet	20.111	0.570	0.449
	ReF-LDM	19.563	0.566	0.347
	RestorerID	19.412	0.559	0.415
	InstantRestore	21.372	0.647	0.323
	FaceMe	20.606	0.616	0.338
	IConFace	18.376	0.570	0.357
CelebHQRef100	DMDNet	22.857	0.648	0.304
	ReF-LDM	21.701	0.615	0.289
	RestorerID	21.857	0.612	0.340
	InstantRestore	22.801	0.674	0.244
	FaceMe	22.598	0.659	0.307
	IConFace	20.479	0.627	0.283

Table 2 shows that reference–GT mismatch is common. The FFHQ-Ref rows are identical because the Moderate and Severe splits share the same GT/reference protocol and differ only in degraded inputs. CelebA-Test-Ref is clearest: 49.11% and 80.77% of Ref_1–GT pairs fall below AdaFace 0.6 and 0.7. This is not a protocol error; reference and GT images are same-identity portraits captured under different pose, expression, lighting, age, makeup, occlusion, or local facial states. A reference-aware result can therefore become more consistent with supplied identity evidence without exactly reproducing the paired target state. The diagnostic panels in Figures 1–4 verify the complementary requirement: IConFace improves identity cues such as facial geometry, eye spacing, nose/mouth shape, and local identity marks, while pose, expression, and layout remain anchored to the degraded target rather than copied from the reference.

Table 3 provides paired target-state checks. IConFace has the lowest PSNR on these benchmarks, consistent with the perception–distortion tradeoff of Blau and Michaeli (2018): a flow/diffusion-style restorer that recovers plausible identity-consistent detail is not optimized for pixel-wise regression to the paired GT. The tradeoff is metric-dependent.

Table 4: GT-AdaFace on the full reference-aware benchmarks. Scores are computed against paired GT targets rather than protocol references.

Method	CelebA	FFHQ-M	FFHQ-S	CHQ100
DMDNet	0.768	0.835	0.170	0.509
ReF-LDM	0.806	0.862	0.661	0.683
RestorerID	0.782	0.828	0.412	0.543
InstantRestore	0.797	0.841	0.551	0.664
FaceMe	0.826	0.894	0.546	0.582
IConFace	0.736	0.814	0.717	0.712

On FFHQ-Ref Severe, IConFace has lower PSNR but better LPIPS than DMDNet and RestorerID and close LPIPS to ReF-LDM. SSIM is competitive on FFHQ-Ref Moderate/Severe and CelebHQRef100, while the CelebA-Test-Ref gap reflects the high Ref_1–GT divergence rate on that benchmark (Table 2), where reference-guided restoration can legitimately deviate from the exact target-state pixel structure. Table 1 gives the analogous paired CelebA-Test blind metrics. Together with the main-paper MUSIQ/CLIP-IQA/MANIQA results, these checks show an identity–perception/distortion tradeoff rather than a PSNR-oriented optimization.

Table 4 makes the GT-based identity picture explicit. The reversal is the key observation: IConFace is not highest on GT-AdaFace in the easier CelebA-Test-Ref and FFHQ-Ref Moderate splits, but becomes strongest on FFHQ-Ref Severe and CelebHQRef100. This directly validates the asymmetric design intent. When the degraded input retains sufficient identity evidence, the reference anchor can introduce mild divergence from the specific target state if Ref_1 and GT differ; conservative baselines may therefore stay closer to the paired GT. When severe degradation destroys that evidence, the reference pathway provides the reliable identity signal, and both Ref-AdaFace and GT-AdaFace improve. Thus high Ref-AdaFace is meaningful only when read together with GT identity, structure checks, and visual evidence that the output keeps the degraded pose/expression rather than copying the reference.

Qualitative material. The following pages expand the qualitative evidence for each benchmark. Reference-aware figures report AdaFace similarity to the first protocol reference image, while no-reference pages omit identity scores because no same-identity reference is supplied. The selected reference–GT diagnostic cases are drawn from the 20 lowest IConFace GT-AdaFace cases in each split and kept when Ref_1 and GT show a visible facial-state gap with clear inter-method differences; the Ref_1 and GT columns expose the actual Ref_1–GT AdaFace values.

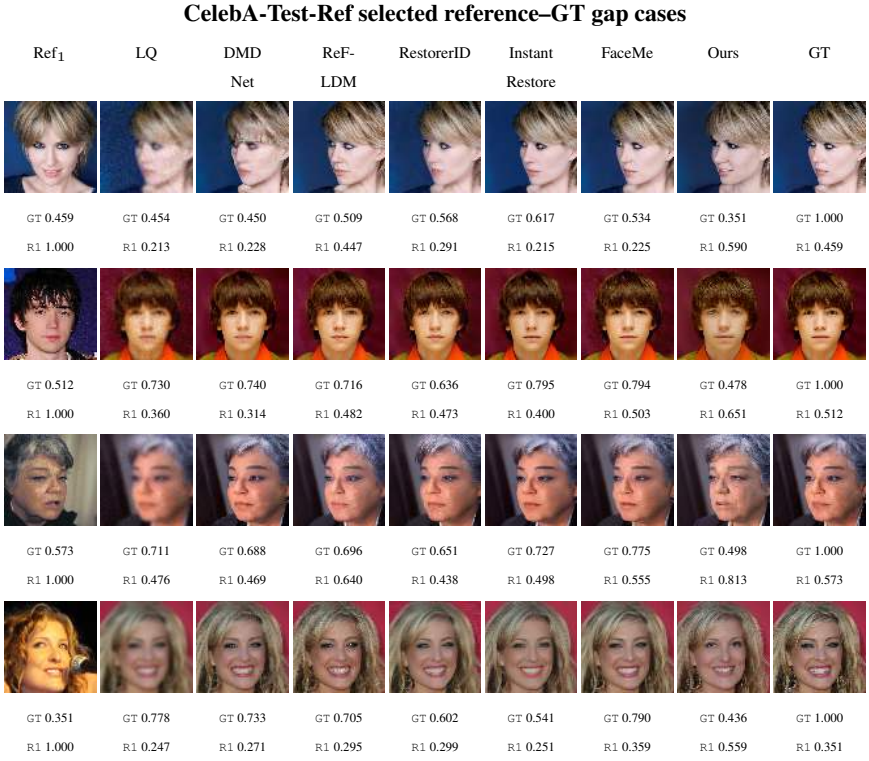

Figure 1: CelebA-Test-Ref reference–GT gap cases. Scores are AdaFace to GT/R1; the final row is an added low Ref_1–GT case. IDs: 18646, 20257, 26060, 02060.

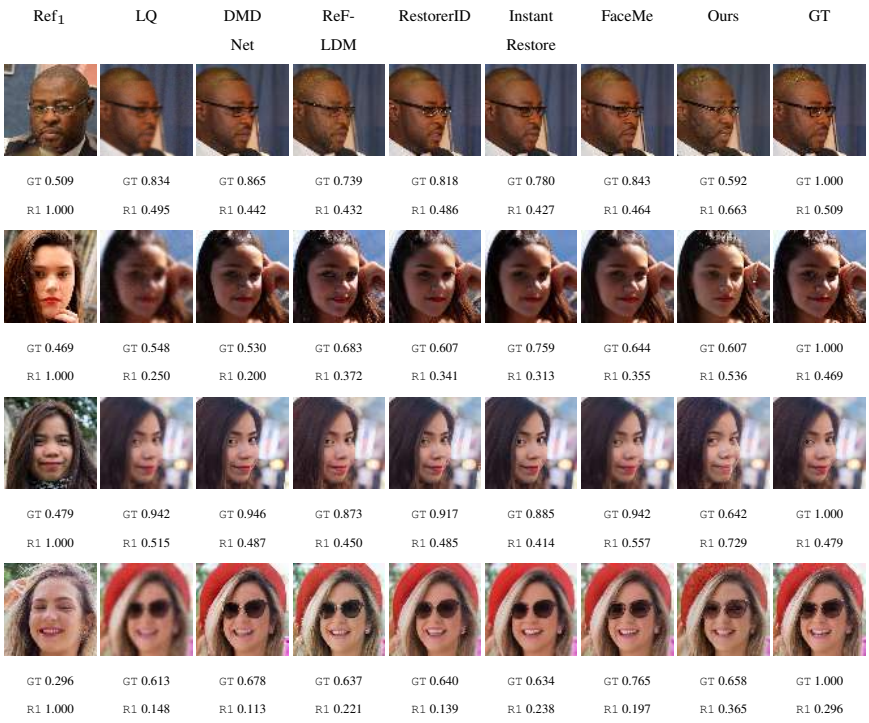

Figure 2: FFHQ-Ref-Moderate reference–GT gap cases. Scores are AdaFace to GT/R1; the final row is an added low Ref_1–GT case. IDs: 57244, 31223, 02293, 37245.

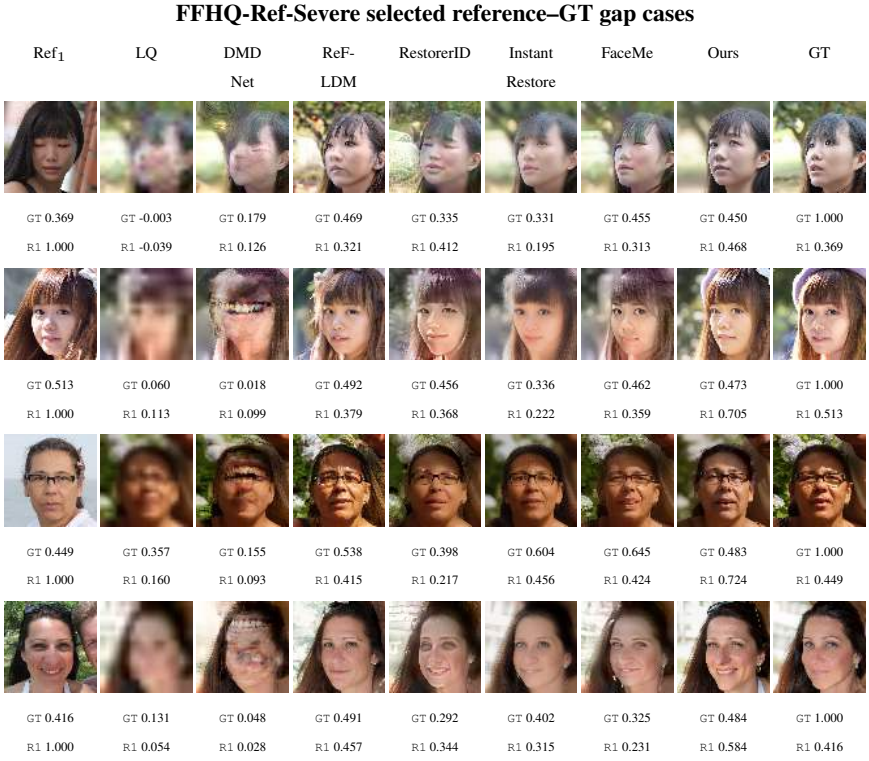

Figure 3: FFHQ-Ref-Severe reference–GT gap cases. Scores are AdaFace to GT/R1; the final row is an added low Ref_1–GT case. IDs: 48700, 04716, 12051, 08081.

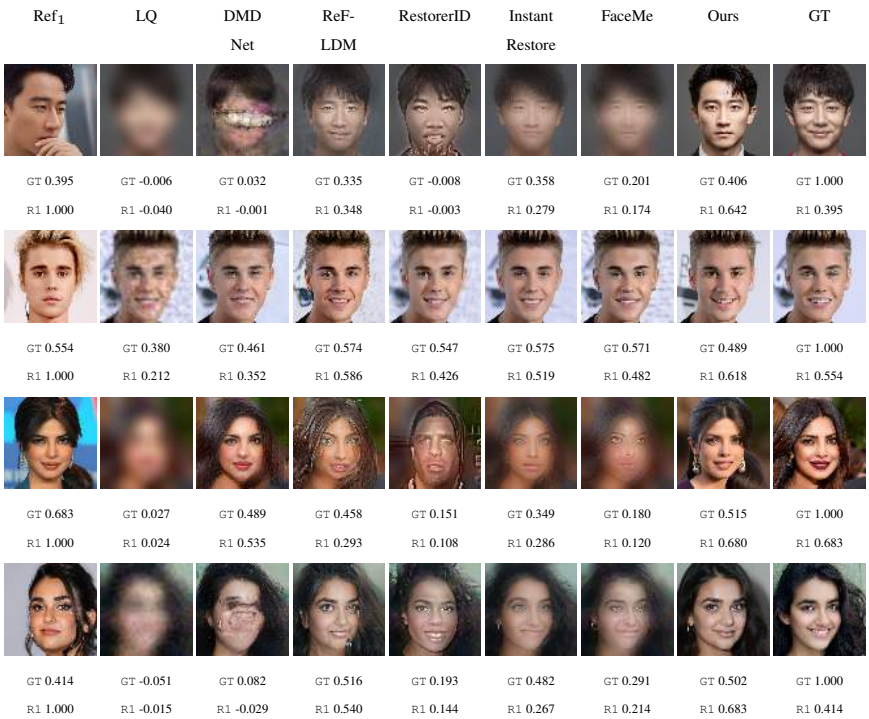

Figure 4: CelebHQRef100 reference–GT gap cases. Scores are AdaFace to GT/R1; the final row is an added low Ref_1–GT case. IDs: 00051__0, 00074__0, 00090__0, 00072__0.

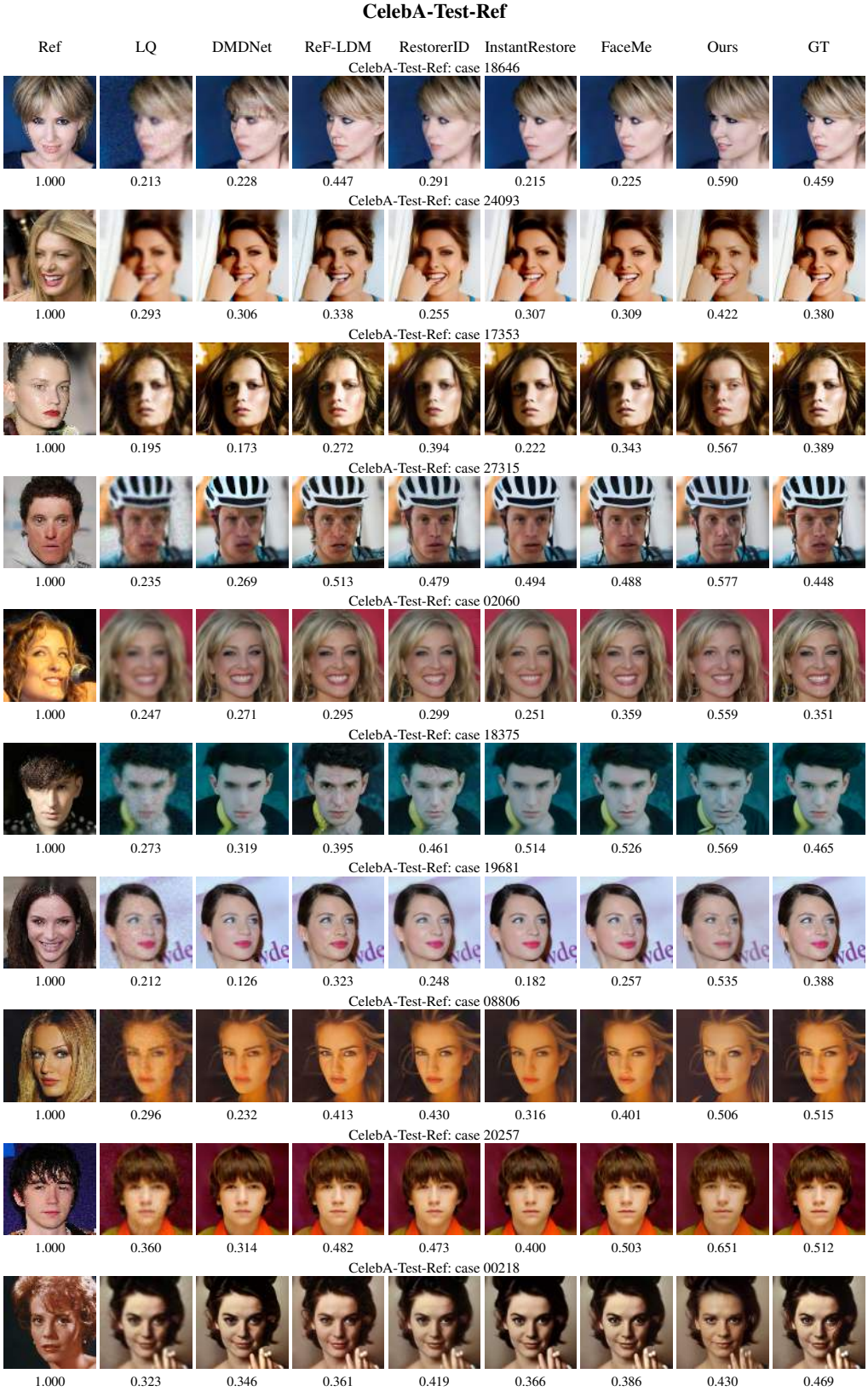

Figure 5: Additional reference-aware qualitative comparisons on CelebA-Test-Ref. Scores under images report AdaFace similarity to the first protocol reference.

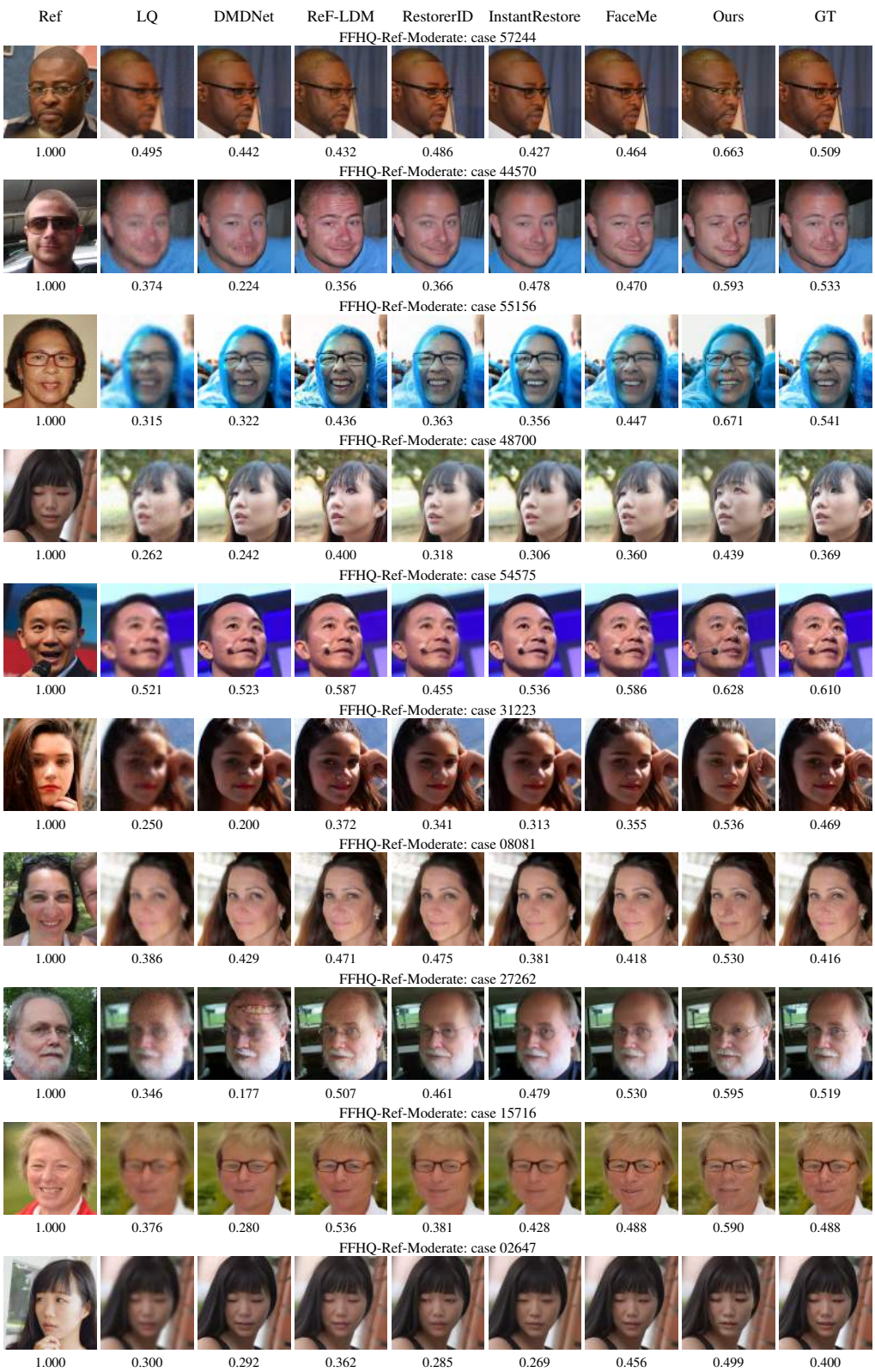

Figure 6: Additional reference-aware qualitative comparisons on FFHQ-Ref Moderate. Scores under images report AdaFace similarity to the first protocol reference.

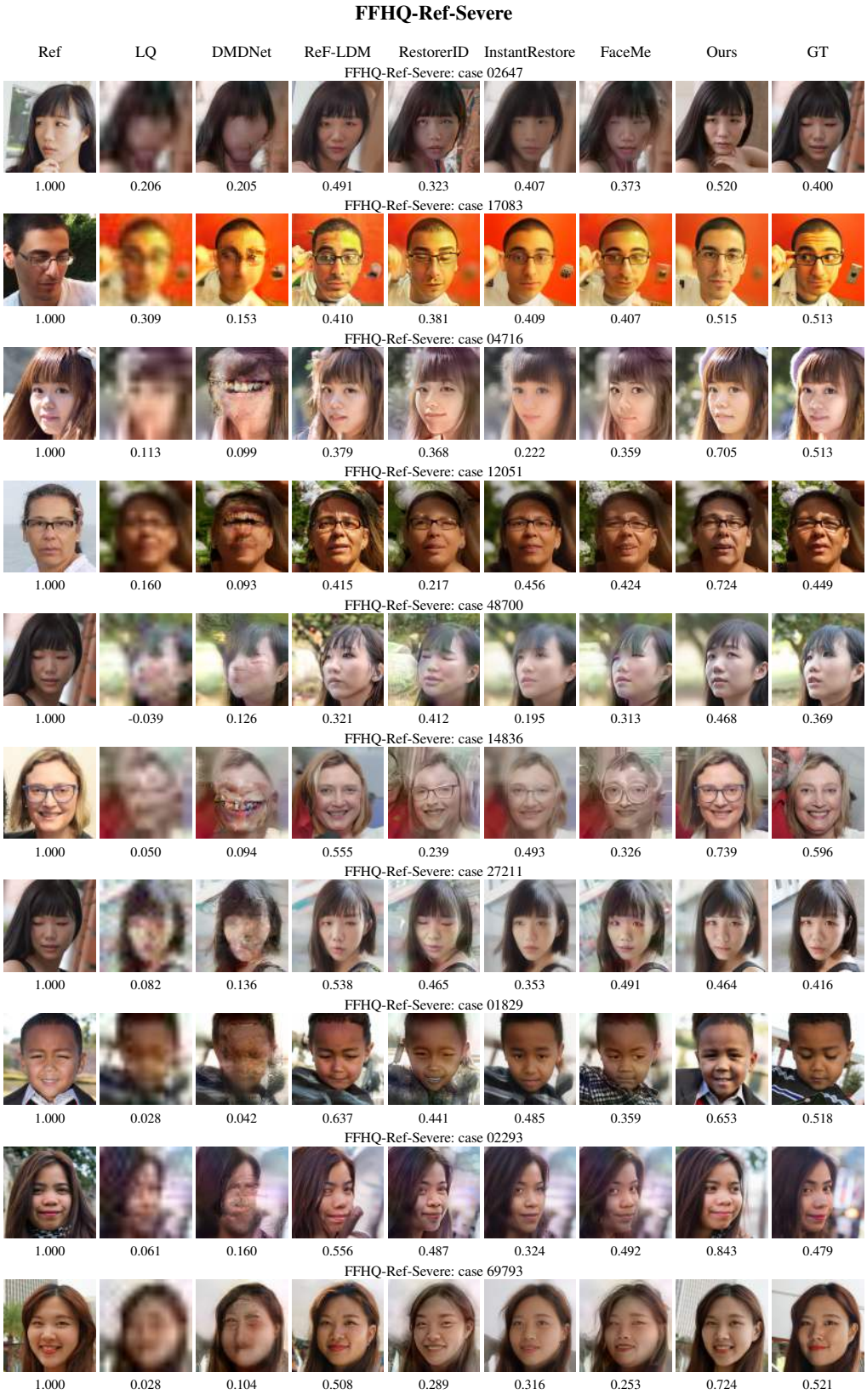

Figure 7: Additional reference-aware qualitative comparisons on FFHQ-Ref Severe. Scores under images report AdaFace similarity to the first protocol reference.

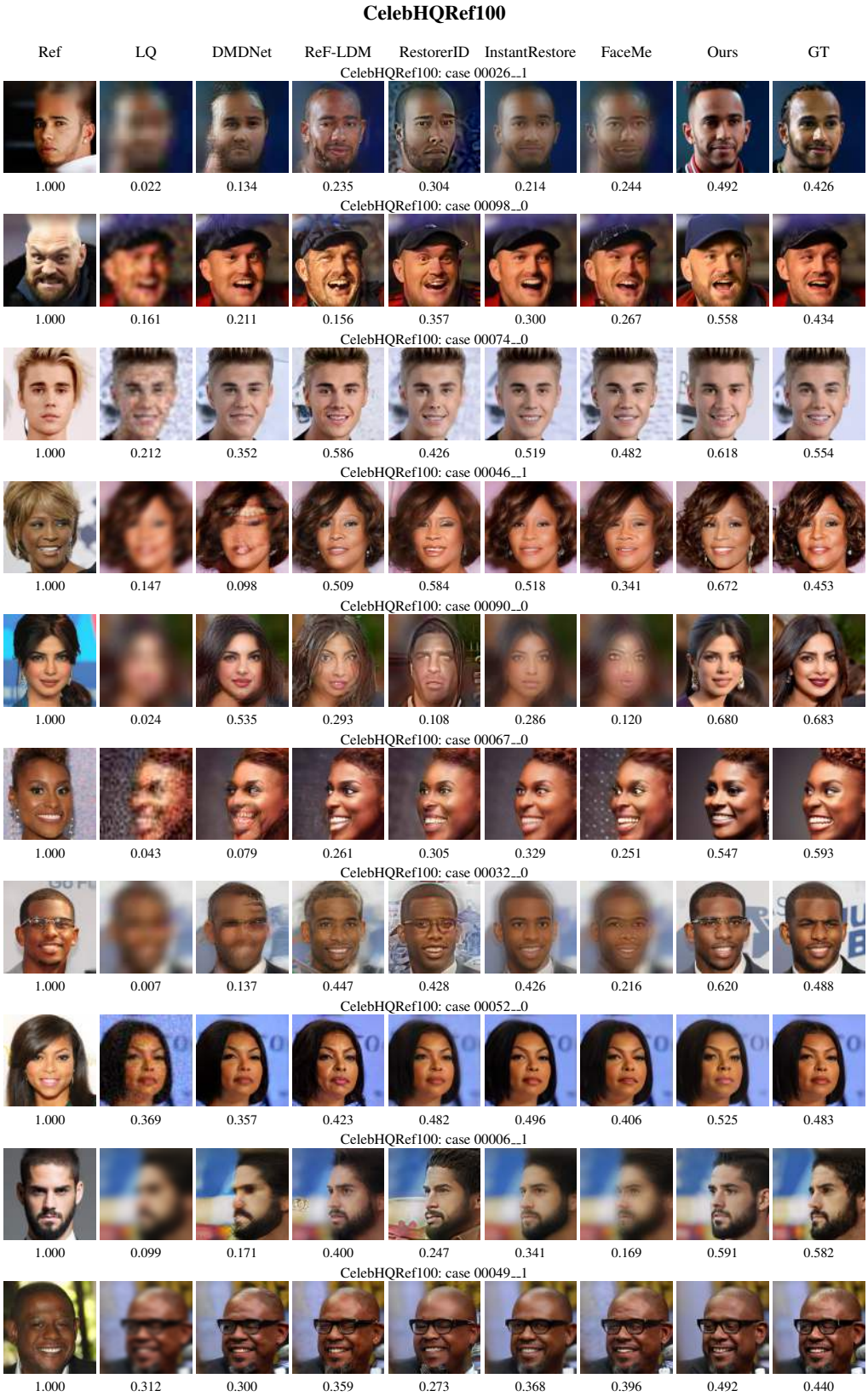

Figure 8: Additional reference-aware qualitative comparisons on CelebHQRef100. Scores under images report AdaFace similarity to the first protocol reference.

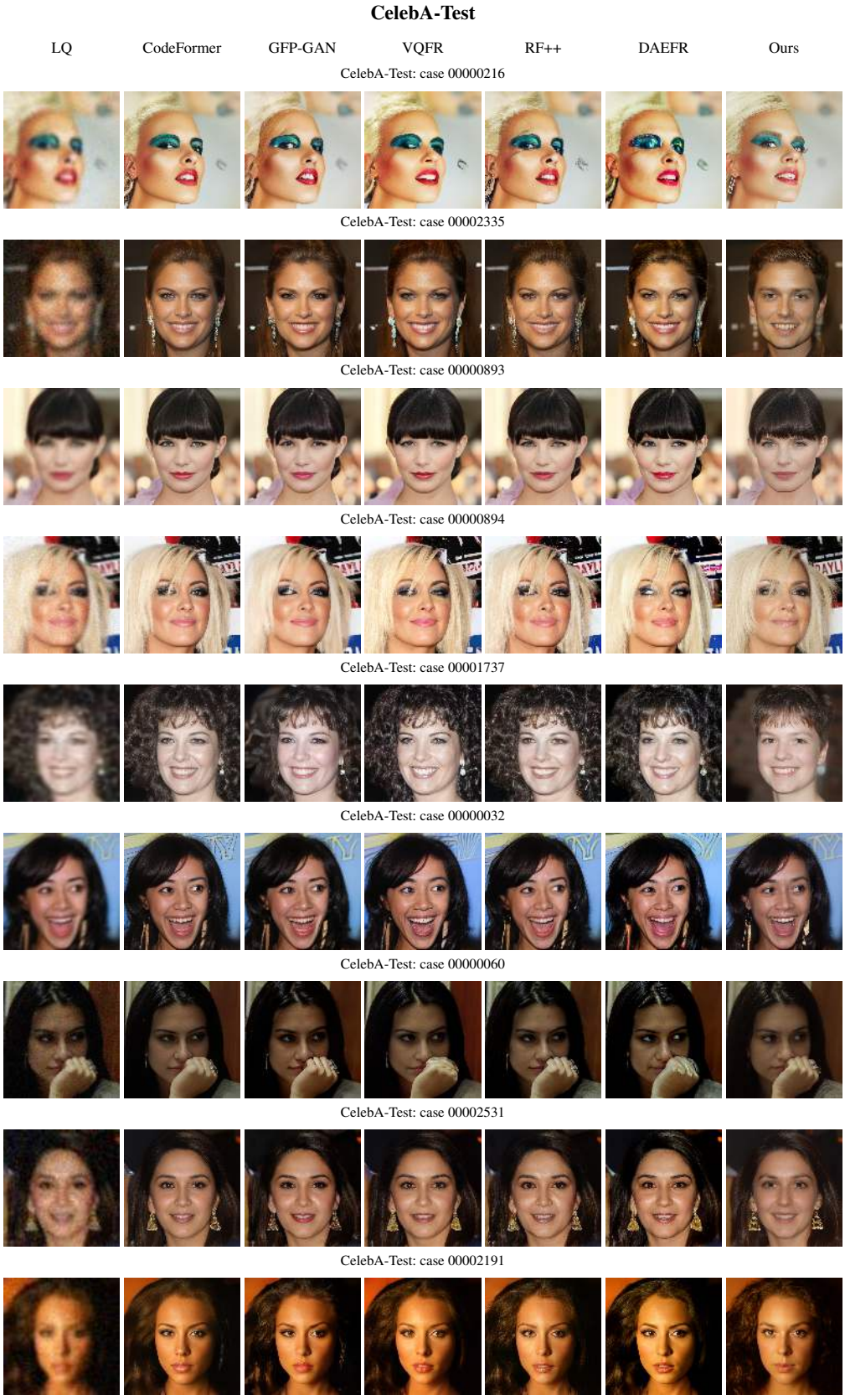

Figure 9: Additional no-reference qualitative comparisons on CelebA-Test.

Figure 10: Additional no-reference qualitative comparisons on LFW.

Figure 11: Additional no-reference qualitative comparisons on CelebChild.

Figure 12: Additional no-reference qualitative comparisons on WebPhoto.

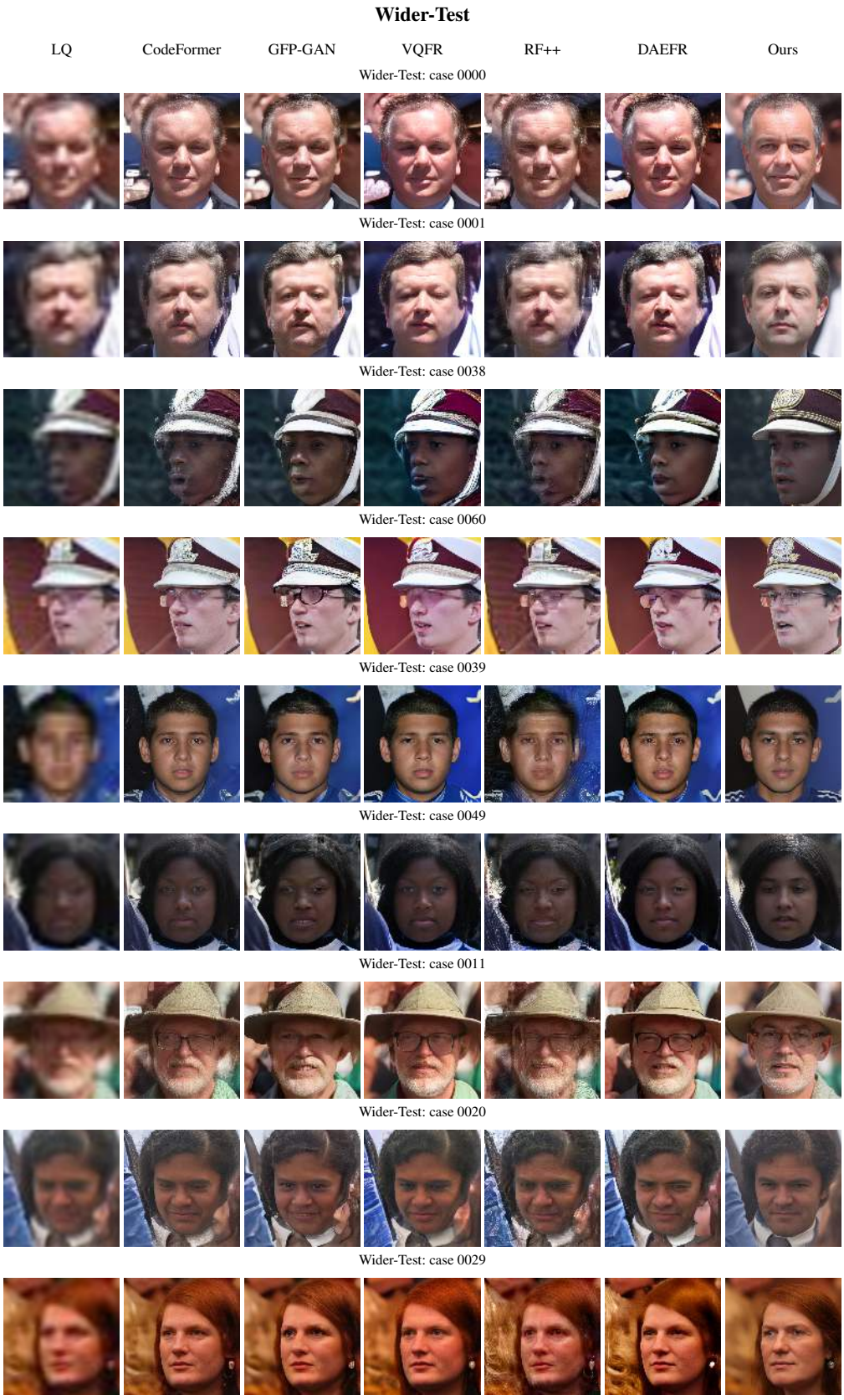

Figure 13: Additional no-reference qualitative comparisons on Wider-Test.

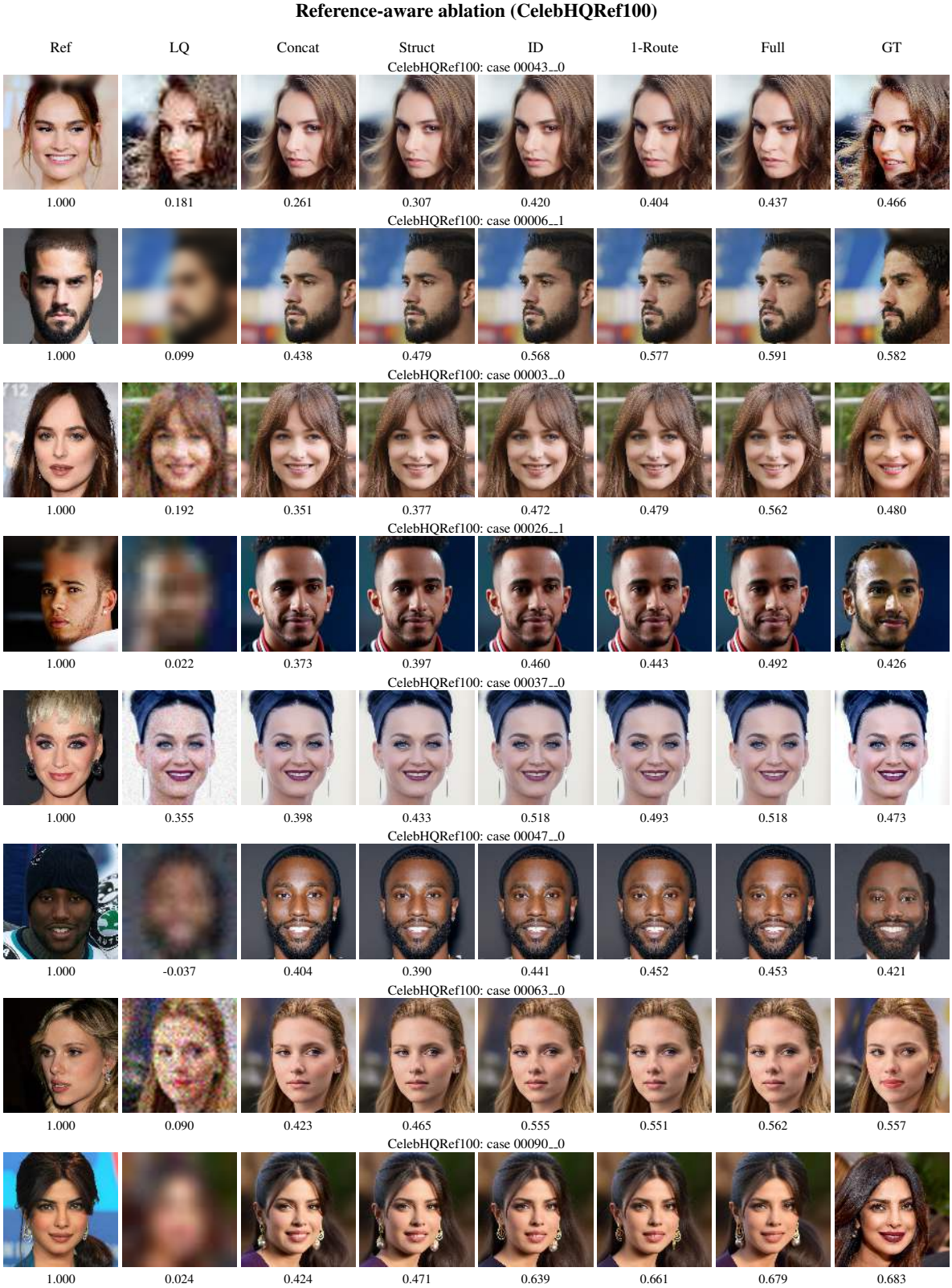

Figure 14: Reference-aware qualitative ablation on CelebHQRef100. Scores under images report AdaFace similarity to the first protocol reference.

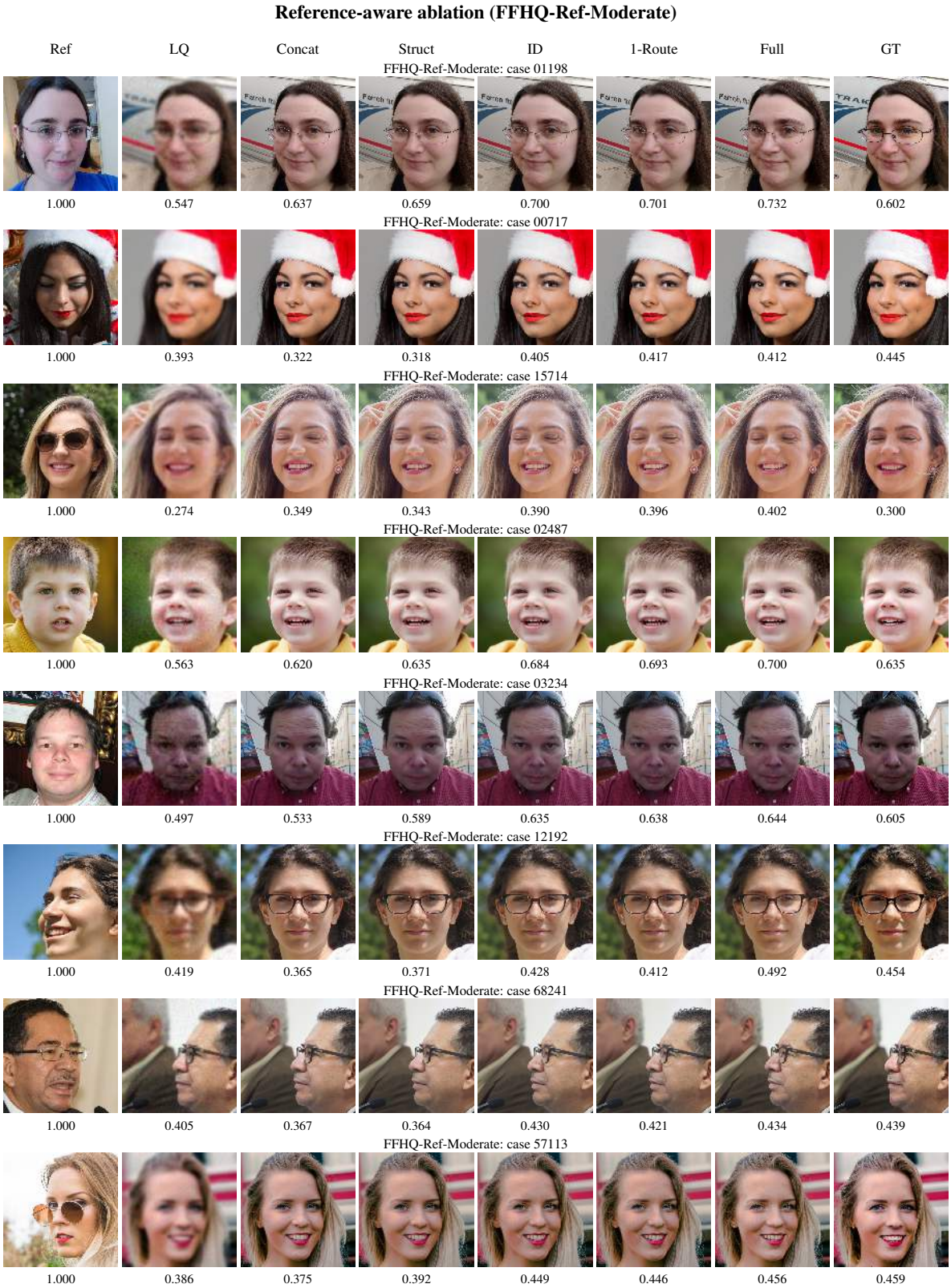

Figure 15: Reference-aware qualitative ablation on FFHQ-Ref Moderate. Scores under images report AdaFace similarity to the first protocol reference.

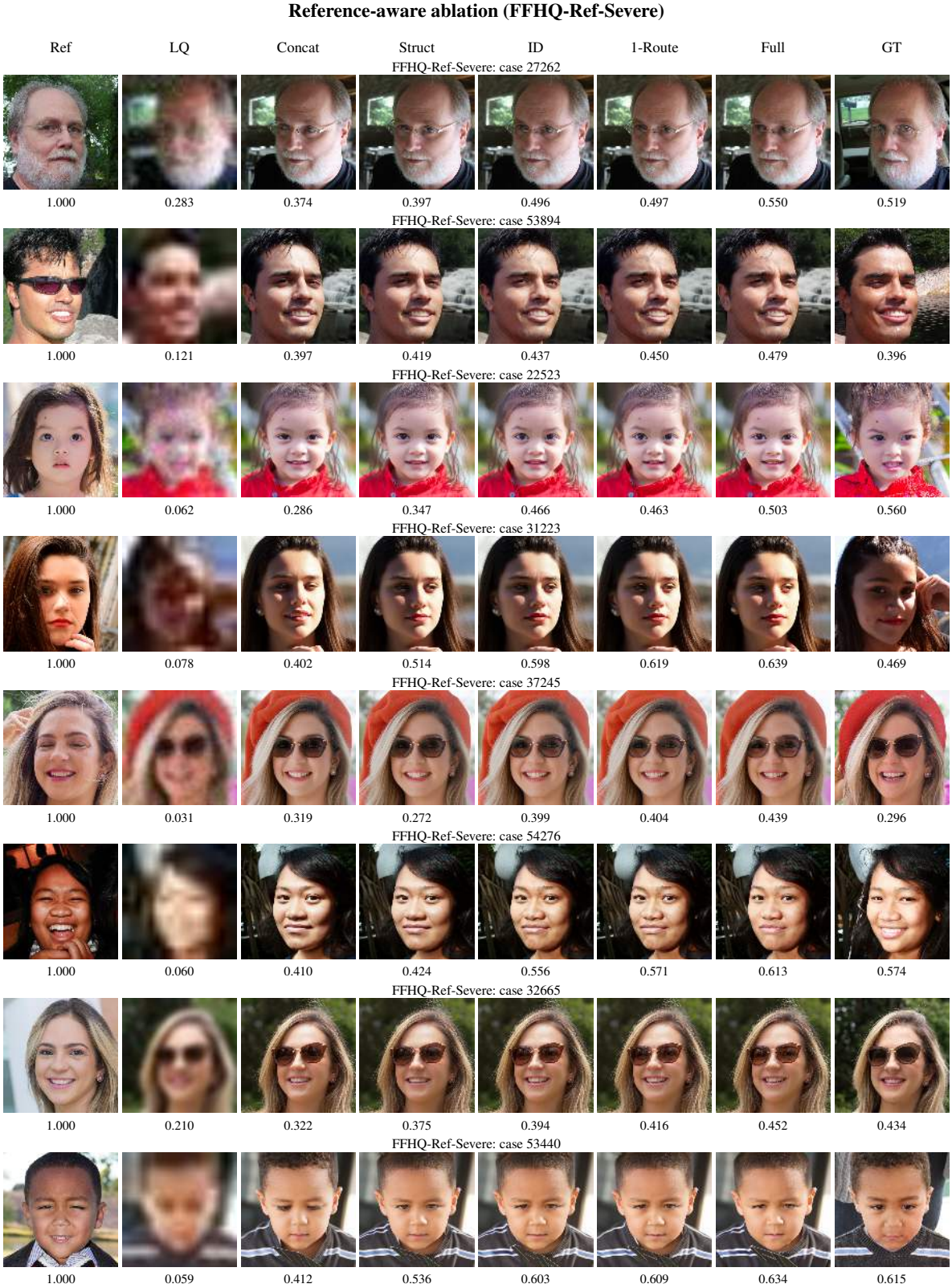

Figure 16: Reference-aware qualitative ablation on FFHQ-Ref Severe. Scores under images report AdaFace similarity to the first protocol reference.