A. Additional evaluations

Figure 8 presents curves for additional metrics. Although the perceptual codecs PICO, HiFiC, C3-WD and CDC substantially outperform the non-perceptual codecs based on human ratings and the perceptually-oriented objective metrics, they do not perform well on PSNR. Conversely, the best-performers on PSNR — DCVC-RT, TCM, ECM, and VVC — perform poorly on perceptual metrics, and require 2-3 times the bitrate to achieve the same perceptual quality as evaluated by viewers.

This further validates the well-known observation that PSNR poorly reflects the human visual system, and that optimizing for it comes in inherent contention with producing reconstructions that humans find to be visually faithful to the originals.

Figures 9 and 10 present objective metric curves, as well as Elo curves from the subjective studies for additional evaluation datasets, Kodak and DIV2K. These showcase that PICO's subjective favorability holds across various datasets.

B. Full model architecture

The architectures of other parts of the model can be found in Figure 11. To derive the hyperparameters of the encoder, neural architecture search was applied in a similar manner to the one of the outer decoder described in the main paper; see Section D for details.

In general, all 3×3 convolutions and ConvScales in the paper are configured to have the number of channels per group be 32, unless stated otherwise.

C. Perceptual training recipe

The training procedure is split into two phases. In the first training phase, we optimize solely for MSE distortion. The learning rate is set to 0.0008 and decayed to 30% and 10% of its initial value at 70% and 90% of training, respectively. In the second phase of the training, we introduce perceptual and GAN losses. The learning rate is decayed to 50%/30%/10% of the initial rate at 30%/60%/80% of training, respectively.

D. Neural architecture search

Here we list the detailed search space and the chosen value for both outer encoder and outer decoder in Table 3. Among the models we ended up with first phase training as described in 4.3, we ranked the models with respect to their compression performance and did a thorough analysis on the impact of different hyperparameters. Take the outer decoder as example, we noticed that under the same runtime budget, putting higher channel number in the low resolution layers (i.e., scale 1) while sacrificing channel number in high resolution layers (i.e., scales 2 and 3) usually gives more benefits

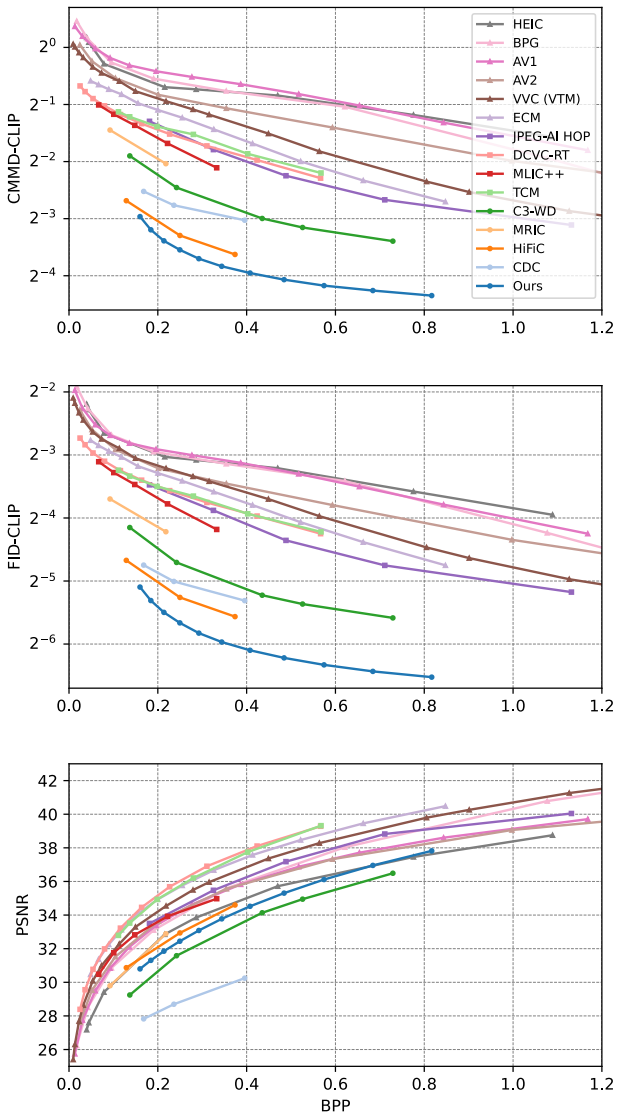


Figure 8. R-D curves for additional metrics.

than other changes (such as repeat nums, 3x3 and 1x1 expansions etc). Note that although the NAS experiments were conducted on iPhone 16 Pro, we cross-validated that the conclusions generalize across different devices, including newer models, like the iPhone 17 Pro on which we reported final runtimes.

E. Baseline codec specifications

BPG [14] encode command:

```
bpgenc <src> \
  -q <qp> \
  -o <enc>
```

The core codec underlying BPG with this distribution uses

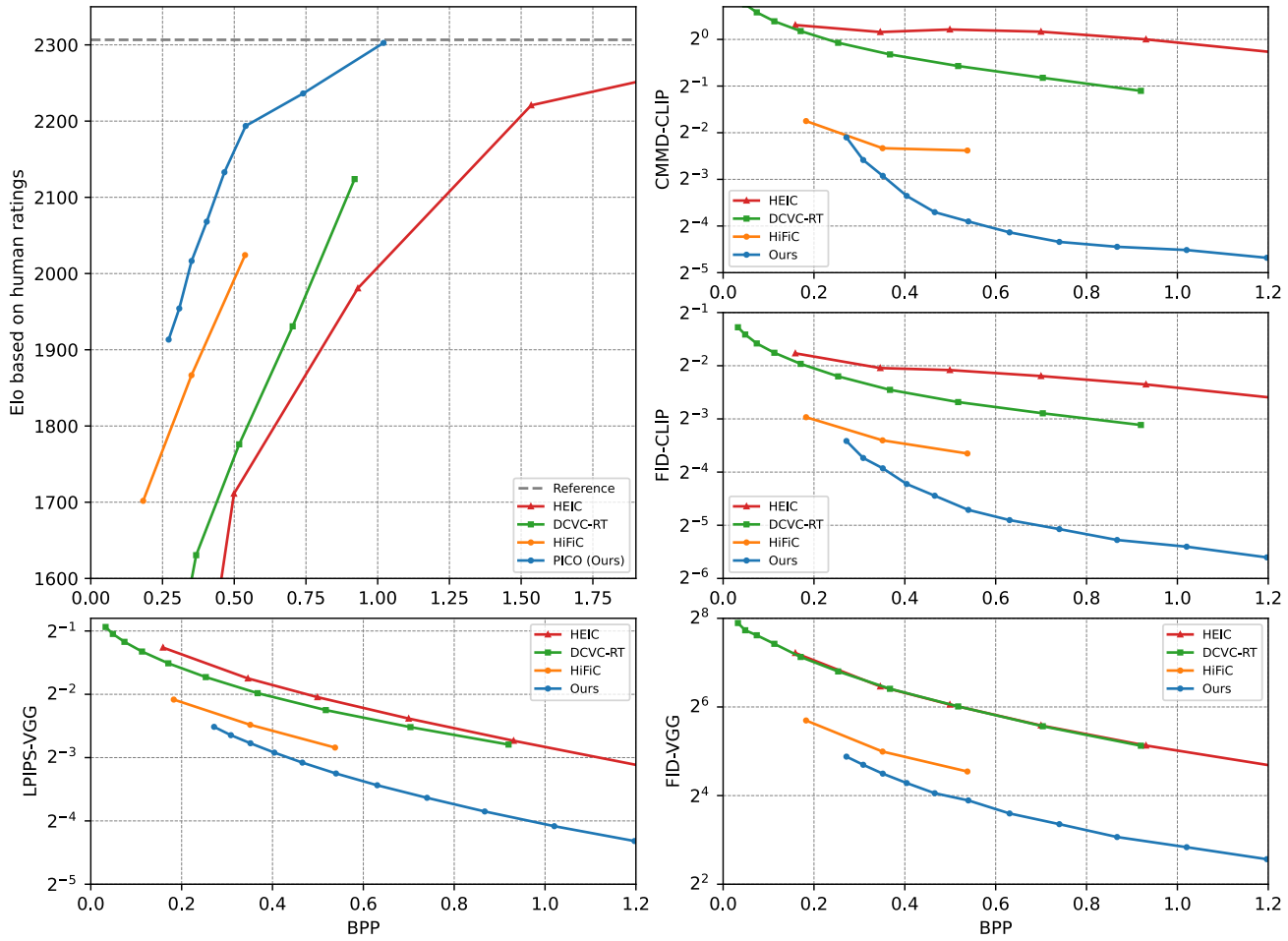


Figure 9. Subjective and objective curves for the Kodak dataset.

x265.
AV1 [16] encode command:

```
aomenc <src> \
  -o <enc> \
  --cq-level=<rate> \
  --end-usage=q \
  --i420
```

AV2 [18] encode command:

```
aomenc <src> \
    -o <enc> \
    --qp=<qp> \
    --psnr \
    --obu \
    --passes=1 \
    --end-usage=q \
    --kf-min-dist=0 \
    --kf-max-dist=0 \
    --use-fixed-qp-offsets=1 \
    --deltaq-mode=0 \
    --enable-tpl-model=0 \
    --cpu-used=8 \
    --enable-keyframe-filtering=0 \
    --i420
```

Note that we benchmarked the AV2 reference implementation: it is the strongest baseline, but is slow and unoptimized (the reference implementations of VVC/ECM were the same or slower).
VVC [15] encode command:

```
EncoderAppStatic -i <src> \
  -c encoder_intra.cfg  \
  -b <enc> \
  -q <qp> \
  --ReconFile /dev/null \
  -fr 1 \
  -f 1 \
  -cf 420
```

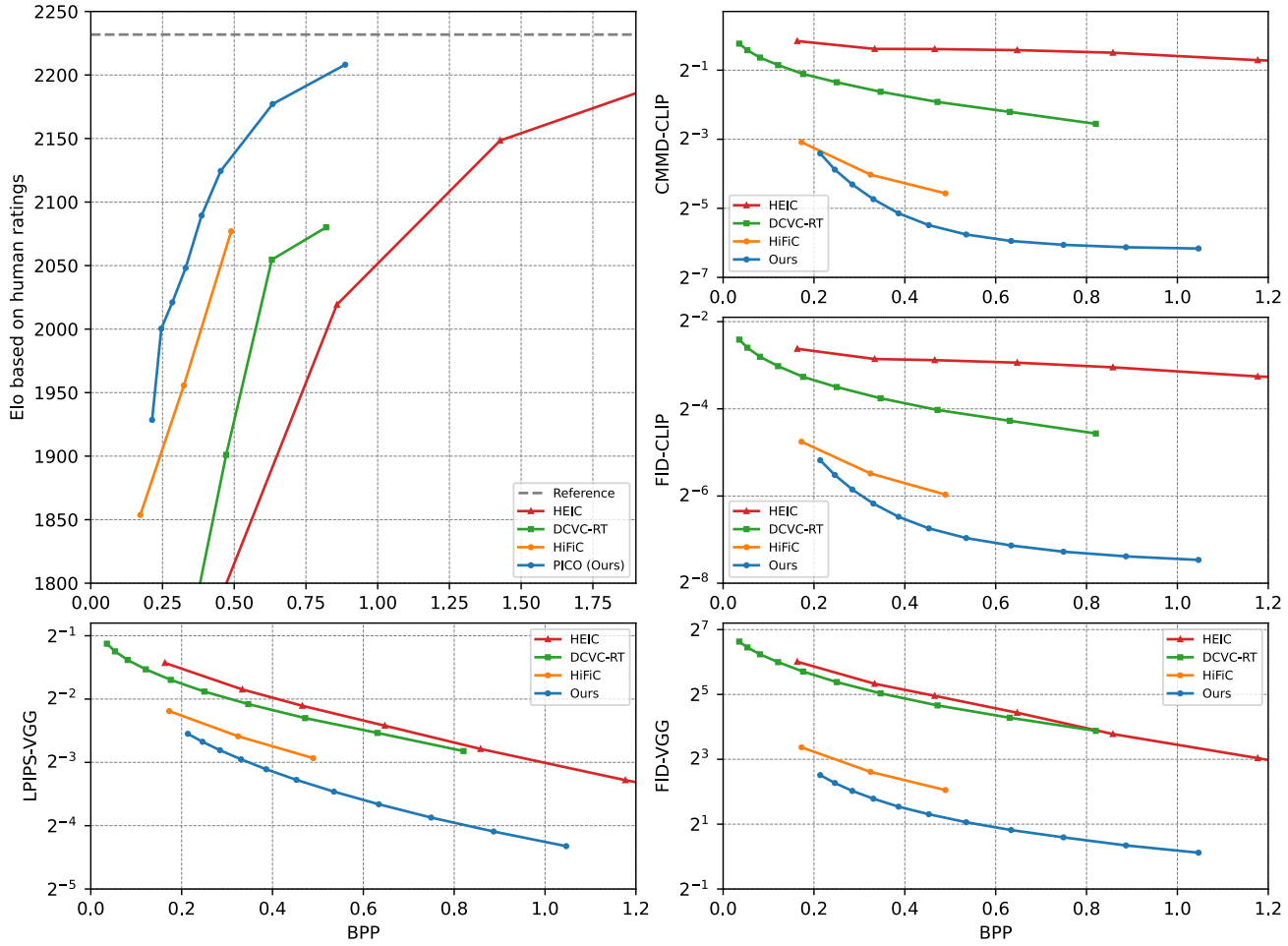


Figure 10. Subjective and objective curves for the DIV2K dataset.

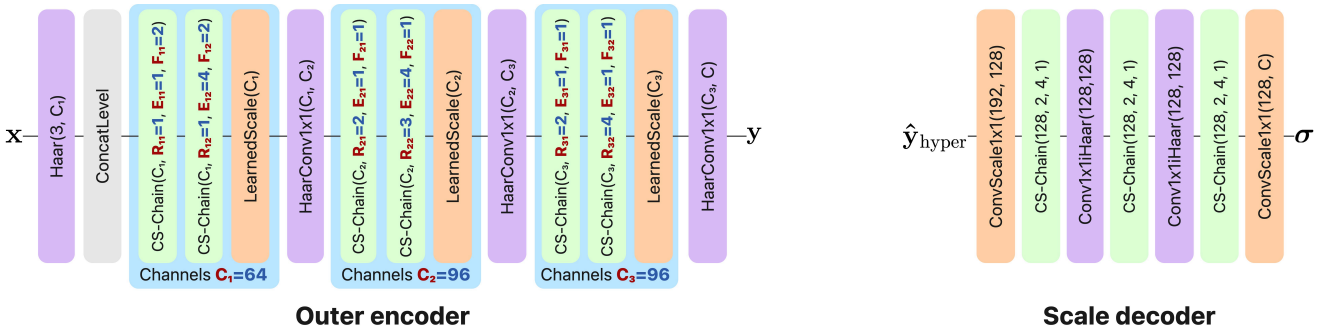


Figure 11. Architectures of the outer encoder and scale decoder. See the main paper body for details.

ECM [17] encode command:

```
EncoderAppStatic -i <src> \
  -c encoder_intra.cfg
  -b <enc> \
  -q <qp> \
  --ReconFile /dev/null \
  -fr 1 \
  -f 1 \
  --CTUSize=256
```

Conversion from RGB to YUV:

```
ffmpeg -y -loglevel quiet -i
"<src>" <pad_option>'-pix_fmt
yuv420p "<dst>"'
```

F. Quality level control

We start with 8 coarse levels $l_c \in 0, \ldots, 7$, which we map to one-hot vectors. We then expand the number of levels to 71, by increasing the level density 10-fold and interpolating the one-hot vector for intermediate levels between the coarse ones. Differently from [5], we apply the level embedding interpolation both during training and inference, rather than as just a post-training step. We condition the encoder and decoder by concatenating to their inputs the interpolated 8-dimensional one-hot tensor broadcasted spatially; we furthermore wrap the latent $\hat{\mathbf{y}}$ with a learned level-conditional channel-wise gain and its inverse. During training, we uniformly sample quality levels, and associate a different Lagrange multiplier λ_l for each. We further add a multiplier α_l reweighing each loss term as function of the level, allowing balancing the gradient during training as it accumulates across different levels. Thus, the total training loss is a combination of the distortion loss D and rate loss R where the latents $\hat{\mathbf{y}}$ and reconstruction $\hat{\mathbf{x}}$ are conditioned on the level:

$$\mathcal{L} = \mathbb{E}_l \left[\alpha_l D\left(\mathbf{x}, \hat{\mathbf{x}}_l\right) + \alpha_l \lambda_l R(\hat{\mathbf{y}}^l_{\text{hyper}}, \hat{\mathbf{y}}^l) \right] \tag{1}$$

G. Subjective study methodology

The subjective study is conducted in a blind pairwise comparison format. Figure 12 shows the interface seen by the human raters. The interface allows for zooming, with default zoom level set to $2x$.

Similar to the CLIC compression challenge [31], the study actively chooses which pair of reconstructions (corresponding to a codec evaluated at a particular rate) are compared against each other using the maximum information gain strategy [49] to maximize comparisons which provide a useful signal. Finally, Bayesian ELO scores are computed based on all the pairwise comparisons.

To avoid noisy voting, Mabyduck performs thorough sanity checking of the reviewers setup with a pre-screening in accordance with the Ishihara color test [50]. The pre-screening checks for color blindness, contrast sensitivity and basic ability to detect compression artifacts. A sample screening study is shown here: https://xp.mabyduck.com/en/latest/pre_screen_image/job/j6ne0x2/.

H. Conv + Haar resampling implementation details

We use Haar wavelets for all resampling operations in the codec — while adding zero additional computation, via a reparametrization trick. In our model, the resampling operation is always coupled with a change of the number of channels. For instance, the encoder might need to downsample by 2× from one spatial scale with C_1 channels, to another with C_2 channels. This could be achieved with applying a Haar transform, followed by a 1×1 convolution mapping from $C_1 \to C_2$. Observing that Haar can be expressed as a simple 4×4 matrix multiplication to the 4 elements of 2×2 spatial blocks, we combine the Haar and the 1x1 convolution into a single 1x1 convolution with a modified weight into which Haar is collapsed, preceded by a factor-2 space-to-depth. The decoder-side conv+iHaar upsampling operation is treated in an analogous way.

I. Limitations

PICO is optimized for perceptual quality specifically for *natural* contents. On extremely simple synthetic contents (*e.g.*, cartoon), PICO uses a higher bitrate compared to conventional codecs to achieve similar quality. This is because the image perfectly fits conventional codecs' autoregressive modeling.

J. Additional reconstructions

Additional visual comparisons of PICO against HiFiC, VVC (VTM) and the original uncompressed image can be found at the end of the supplementary materials.

Multiple issues can be seen in HiFiC relative to PICO:

- Over-synthesis: it hallucinates details, at the cost of fidelity to the original image.
- Synthesis of incorrect statistics relative to the original: it introduces patterns to smooth surfaces, and over-sharpens edges and textures.
- HiFiC is often unable to keep small text legible.
- HiFiC exhibits noticeable structured repetitive patterns, where the underlying texture is more random.

Hyperparameter			Search Space	Final value
Scale 1	channels	C_1	[32, 64]	64
	1st CS-Chain	R_{11}	[1, 2]	1
		E_{11}	[1]	1
		F_{11}	[1, 2]	2
	2nd CS-Chain	R_{12}	[1, 2]	1
		E_{12}	[1, 2, 4]	4
		F_{12}	[1, 2]	2
Scale 2	channels	C_2	[64, 96]	96
	1st CS-Chain	R_{21}	[2, 4]	2
		E_2	[1]	1
		F_{21}	[1]	1
	2nd CS-Chain	R_{22}	[1, 2, 3]	3
		E_{22}	[1, 2, 4]	4
		F_{22}	[1, 2]	1
Scale 3	channels	C_3	[96, 128, 160]	96
	1st CS-Chain	R_{31}	[2, 4, 6]	2
		E_{31}	[1]	1
		F_{31}	[1]	1
	2nd CS-Chain	R_{32}	[2, 4]	4
		E_{32}	[1, 2, 4]	1
		F_{32}	[1, 2]	1

(a) Outer Encoder

Hyperparameter			Search Space	Final value
Scale 1	channels	C_1	[96, 128, 160]	160
	1st CS-Chain	R_{11}	[2, 3, 4]	3
		E_{11}	[1]	1
		F_{11}	[1]	1
	2nd CS-Chain	R_{12}	[2, 3]	2
		E_{12}	[1, 2, 4]	1
		F_{12}	[1, 2]	2
Scale 2	channels	C_2	[64, 96]	64
	1st CS-Chain	R_{21}	[1, 2, 3]	2
		E_2	[1]	1
		F_{21}	[1]	1
	2nd CS-Chain	R_{22}	[1, 2]	1
		E_{22}	[1, 2, 4]	4
		F_{22}	[1, 2]	2
Scale 3	channels	C_3	[32, 64]	32
	1st CS-Chain	R_{31}	[1, 2]	2
		E_{31}	[1, 2]	2
		F_{31}	[1, 2]	2
	2nd CS-Chain	R_{32}	[1, 2]	1
		E_{32}	[1, 2, 3]	3
		F_{32}	[1, 2]	2

(b) Outer Decoder

Table 3. Neural architecture search summary for outer encoder and decoder.

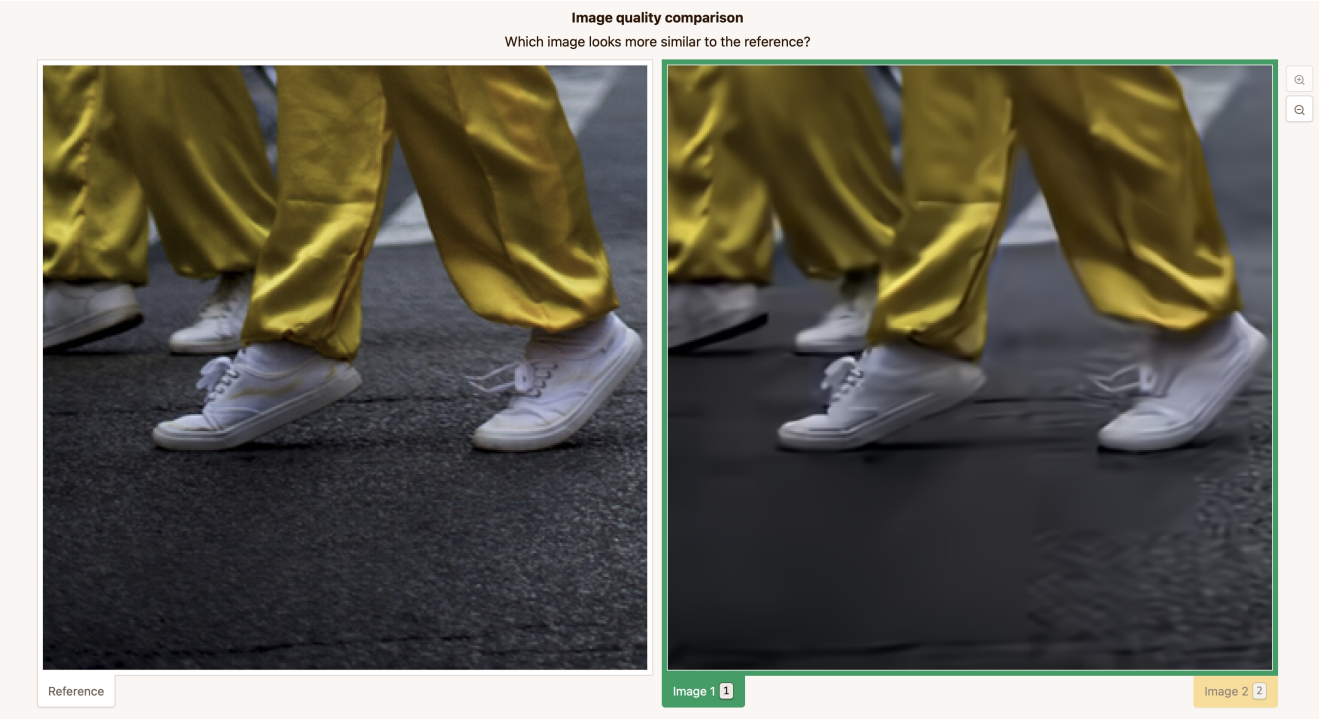
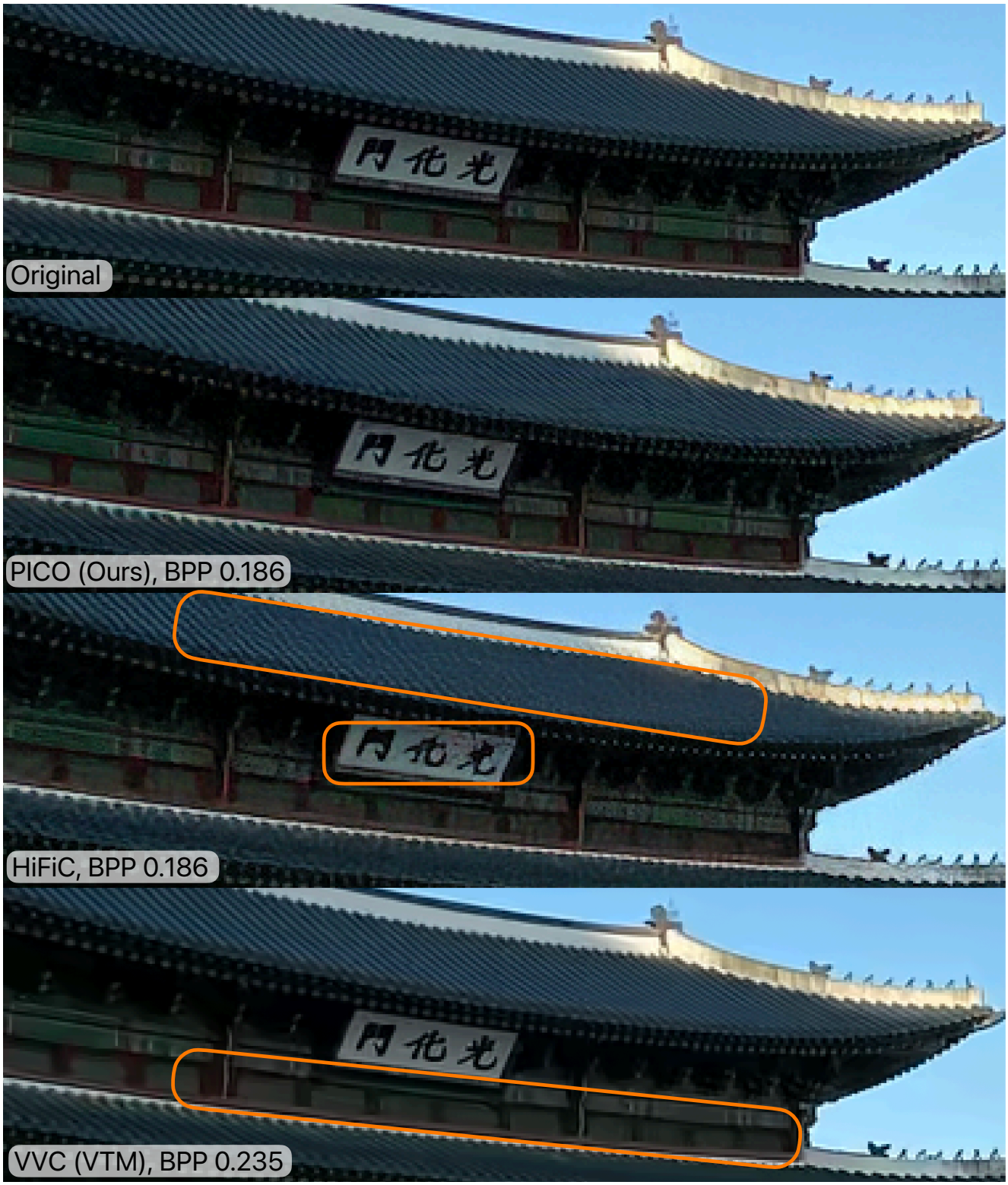


Figure 12. Screenshot of the subjective study interface as seen by the human raters

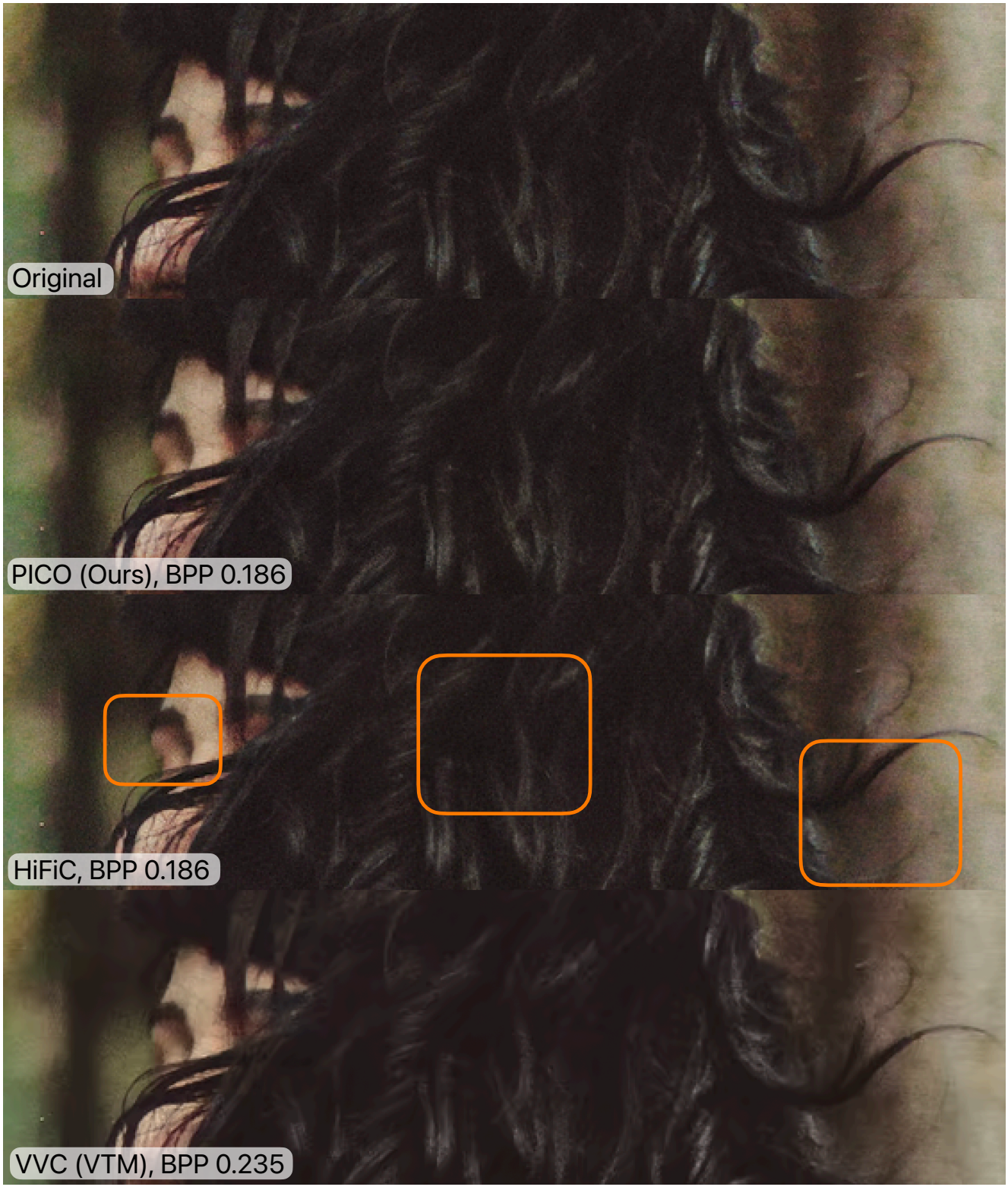

Original
PICO (Ours), BPP 0.341
HiFiC, BPP 0.346
VVC (VTM), BPP 0.360


13 Ropsten · Norsborg
14 Mörby centrum · Fruängen
17 Åkeshov · Skarpnäck
18 Alvik · Farsta strand
19 Hässelby strand · Hagsätra
Original
PICO (Ours), BPP 0.353
HiFiC, BPP 0.353
VVC (VTM), BPP 0.370


Original
PICO (Ours), BPP 0.186
HiFiC, BPP 0.186
VVC (VTM), BPP 0.235


Clarenville
Come By Chance
Sunnyside
Original
Clarenville
Come By Chance
Sunnyside
PICO (Ours), BPP 0.095
Clarenville
Come By Chance
Sunnyside
HiFiC, BPP 0.097
Clarenville
Come By Chance
Sunnyside
VVC (VTM), BPP 0.105


Original
PICO (Ours), BPP 0.186
HiFiC, BPP 0.186
VVC (VTM), BPP 0.235


Original
PICO (Ours), BPP 0.186
HiFiC, BPP 0.186
VVC (VTM), BPP 0.235


Original
PICO (Ours), BPP 0.273
HiFiC, BPP 0.273
VVC (VTM), BPP 0.314