
Supplemental Materials

Anonymous Author(s)

Affiliation

Address

email

1 Anonymized Resources

2 We provide an anonymous project website, which includes a demo video showcasing our method.
3 This website can be found at <https://anonymous.4open.science/w/ESCA-5FC8/>.

4 A Proof of Scaling Invariance

5 Let the convolution layer take an input tensor $X \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ and a weight tensor $W \in$
6 $\mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k_h \times k_w}$. For clarity we fix a single output channel and suppress the output-channel index;
7 the argument is identical for every output filter. With the usual definition of discrete convolution (*),
8 the pre-activation output is

$$Y = \sum_{c=1}^{C_{\text{in}}} W[c] * X[c], \quad (1)$$

9 where $W[c] \in \mathbb{R}^{k_h \times k_w}$ and $X[c] \in \mathbb{R}^{H \times W}$ denote the c -th input-channel kernel and feature map,
10 respectively.

11 Choose positive scalars $s_1, \dots, s_{C_{\text{in}}}$. Define the scaled activations and the compensated weights

$$\tilde{X}[c] = s_c X[c], \quad \tilde{W}[c] = \frac{1}{s_c} W[c], \quad c = 1, \dots, C_{\text{in}} \quad (2)$$

12 For any scalar $\alpha \in \mathbb{R}$ and tensors A, B of compatible shape, convolution is bilinear:

$$(\alpha A) * B = \alpha (A * B), \quad A * (\alpha B) = \alpha (A * B) \quad (3)$$

13 Using Equation 3 with $\alpha = s_c$ and $\alpha = 1/s_c$,

$$\left(\frac{1}{s_c} W[c]\right) * (s_c X[c]) = \frac{1}{s_c} s_c (W[c] * X[c]) = W[c] * X[c] \quad (4)$$

14 Summing Equation 4 over all input channels reproduces Equation 1:

$$\tilde{Y} = \sum_{c=1}^{C_{\text{in}}} \tilde{W}[c] * \tilde{X}[c] = \sum_{c=1}^{C_{\text{in}}} W[c] * X[c] = Y \quad (5)$$

15 Channel-wise scaling of activations, paired with the reciprocal scaling of the corresponding kernels,
16 leaves the convolution output unchanged. Hence $\tilde{Y} = Y$, proving the scale invariance claim.

17 B Proof of Fusing Scaling into Previous Layer Weights

18 Let

$$X^{(1)} \in \mathbb{R}^{C_{in} \times H \times W}, \quad W^{(1)} \in \mathbb{R}^{C_{out} \times C_{in} \times k_h \times k_w}, \quad B^{(1)} \in \mathbb{R}^{C_{out}} \quad (6)$$

19 and define the

$$X^{(2)} = W^{(1)} * X^{(1)} + B^{(1)} \quad (7)$$

20 Let $s \in \mathbb{R}_{>0}^{C_{out}}$ be a per-output-channel scale, and write $s \otimes T$ for broadcast Hadamard multiplication
21 along all trailing dimensions of a tensor T whose first index has size C_{out} .

Claim.

$$\tilde{X}^{(2)} = s \otimes X^{(2)}[:, :, :] = (s \otimes W^{(1)}[:, :, :]) * X^{(1)} + s \otimes B^{(1)} \quad (8)$$

22 Fix an output channel $c \in \{1, \dots, C_{out}\}$ and spatial index (i, j) . By definition of convolution with
23 bias,

$$X_{c,i,j}^{(2)} = \sum_{m=1}^{C_{in}} \sum_{u=0}^{k_h-1} \sum_{v=0}^{k_w-1} W_{c,m,u,v}^{(1)} X_{m,i-u,j-v}^{(1)} + B_c^{(1)} \quad (9)$$

24 Multiply Equation 9 by the scalar s_c :

$$s_c X_{c,i,j}^{(2)} = \sum_{m,u,v} (s_c W_{c,m,u,v}^{(1)}) X_{m,i-u,j-v}^{(1)} + s_c B_c^{(1)} \quad (10)$$

25 The summation term in Equation 10 is exactly the (c, i, j) -entry of the convolution $(s \otimes W^{(1)}) * X^{(1)}$,
26 while the final term is the (c, i, j) -entry of $s \otimes B^{(1)}$ (broadcast spatially). Since (2) holds for every
27 c, i, j , we obtain

$$s \otimes X^{(2)} = (s \otimes W^{(1)}) * X^{(1)} + s \otimes B^{(1)} \quad (11)$$

28 which proves the claim.

29 Scaling each output channel by s can be equivalently implemented by scaling the corresponding
30 output-channel kernels in $W^{(1)}$ before convolution. In quantization or inference-time fusion, this
31 lets us absorb channel-wise activation rescaling into the layer's weights, avoiding an extra runtime
32 operation.

33 C Proof of scaling invariance before and after im2col

34 Let $X \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$ be the activation tensor feeding a ConvTranspose2d layer. $\text{im2col}(\cdot)$ convert
35 its argument to a 2-D matrix in which every column is the receptive-field patch that contributes to
36 one output location.

$$X_{col} = \text{im2col}(X) \in \mathbb{R}^{(C_{in} K_h K_w) \times N} \quad (12)$$

37 where $N = H_{out} W_{out}$ is the number of spatial sites produced by the layer.

38 $W \in \mathbb{R}^{C_{in} \times C_{out} \times K_h \times K_w}$ be the kernel of the transposed convolution, reshaped into

$$W_{mat} = \text{reshape}(W) \in \mathbb{R}^{(C_{in} K_h K_w) \times (C_{out})} \quad (13)$$

39 After im2col , transposed convolution is a plain matrix multiply followed by col2im accumulation

$$Y_{col} = W_{mat}^T X_{col} \quad (14)$$

40 then $Y = \text{col2im}(Y_{col})$.

41 Let the per-channel scale vector be $s = [s_1, \dots, s_{C_{in}}]^T$ with $s_c > 0$. Define $S_{act} = \text{diag}(s) \otimes I_{K_h K_w}$,
42 where $I_{K_h K_w}$ is the identity matrix of size $K_h \times K_w$. The Kronecker product \otimes constructs a block-
43 diagonal matrix S_{act} , and $S_{act} \in \mathbb{R}^{(C_{in} K_h K_w) \times (C_{in} K_h K_w)}$. Every column block belonging to
44 channel c is multiplied by the same scalar s_c .

45 We scale activations and invert-scale the weights $\tilde{X}_{col} = S_{act} X_{col}$ and $\tilde{W}_{mat} = S_{act}^{-1} W_{mat}$.

46 Propagating the scaled quantities through the same algebra as (14).

$$\tilde{Y}_{col} = \tilde{W}_{mat}^T \tilde{X}_{col} = (S_{act}^{-1} W_{mat})^T (S_{act} X_{col}) \quad (15)$$

47 From S_{act} is diagonal, $S_{act}^{-1} = S_{act}^{-1T}$.

$$\tilde{Y}_{col} = W_{mat}^T S_{act}^{-1} S_{act} X_{col} = W_{mat}^T X_{col} = Y \quad (16)$$

48 D Sensitivity to Facial-Feature-Aware Threshold k (4-bit)

49 We test the facial-feature-aware smoothing threshold k from 30 to 95 for our full 4-bit ESCA pipeline
 50 (ICAS + FFAS + UV-weighted Hessian PTQ). Figure 1 and Table 1 report perceptual quality on
 51 the front, left, and right camera views. Quality improves monotonically as k increases from 30
 52 to the 75–80 band, after which it degrades. The optimal window ($k \approx 75$ –80) yields the highest
 FovVideoVDP on all three views, so we adopt $k = 75$ as the default throughout the main paper.

Table 1: 4-bit reconstruction quality (VDP \uparrow) vs. smoothing threshold k . Rows are camera views; columns list the k values.

Metric	View	30	40	50	60	70	75	80	85	90	95
FovVideoVDP \uparrow	Front	5.6446	5.6559	5.6575	5.6681	5.6925	5.7184	5.8541	5.8274	5.8230	5.7862
	Right	4.9189	4.9273	4.9394	4.9427	4.9589	4.9605	4.9550	4.9436	4.8819	4.8187
	Left	4.9074	4.9145	4.9167	4.9337	4.9683	4.9795	4.9660	4.9659	4.9146	4.8495

53
 54 For small k values, Our model employs excessive smoothing, yielding FovVideoVDP scores compa-
 55 rable to ICAS-UV due to suppressed high-frequency facial details and minimal improvement. As
 56 k increases to mid-range values, smoothing intensity decreases and VDP scores rise monotonically
 57 across all viewpoints, indicating progressive fine-detail recovery. Optimal performance occurs within
 58 the narrow range $k = 75/80k$, achieving balanced outlier suppression and detail preservation. For
 59 $k \geq 85$, FFAS provides negligible channel scaling, causing performance degradation toward the
 60 no-smoothing baseline (UV-weighted PTQ without ICAS).

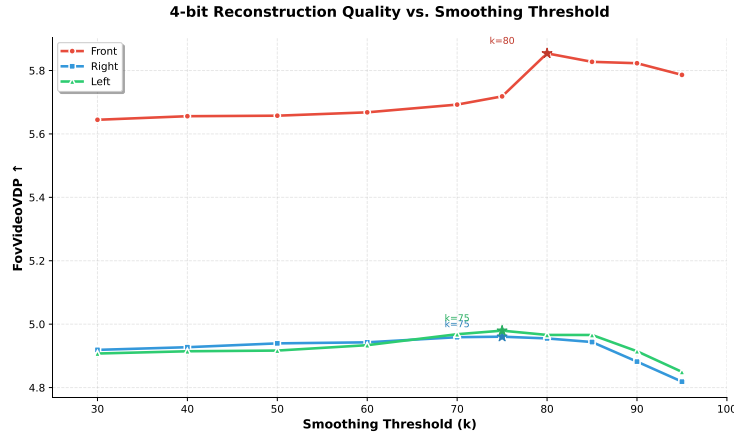


Figure 1: Trend of FovVideoVDP vs. k (averaged over views). Peak quality occurs in the $k = 75$ –80 region; quality drops for $k > 80$.

61 E Additional Rendering Results



(a) Front



(b) Left



(c) Right

(a) Full-precision baseline (FP32)



(d) Front



(e) Left



(f) Right

(b) Best prior low-bit baseline (4-bit)



(g) Front



(h) Left



(i) Right

(c) ESCA (4-bit, ours)

Figure 2: **Qualitative comparison of avatar reconstructions.** Each row shows renderings produced by a different model (full-precision, prior 4-bit baseline, and our ESCA, FFAS+UV). Columns correspond to three viewpoints. ESCA maintains facial detail (eyes, mouth, skin texture) that the prior low-bit model blurs, while with high fidelity.

