CLIPtone-GO: Geometry-Aware, Gradient-Orthogonalized Text-Guided Tone Mapping

Supplementary File

Input **CLIPtone [2]**

CLIPtone-GO (Ours)

Figure 1. Extended View of Figure 1.

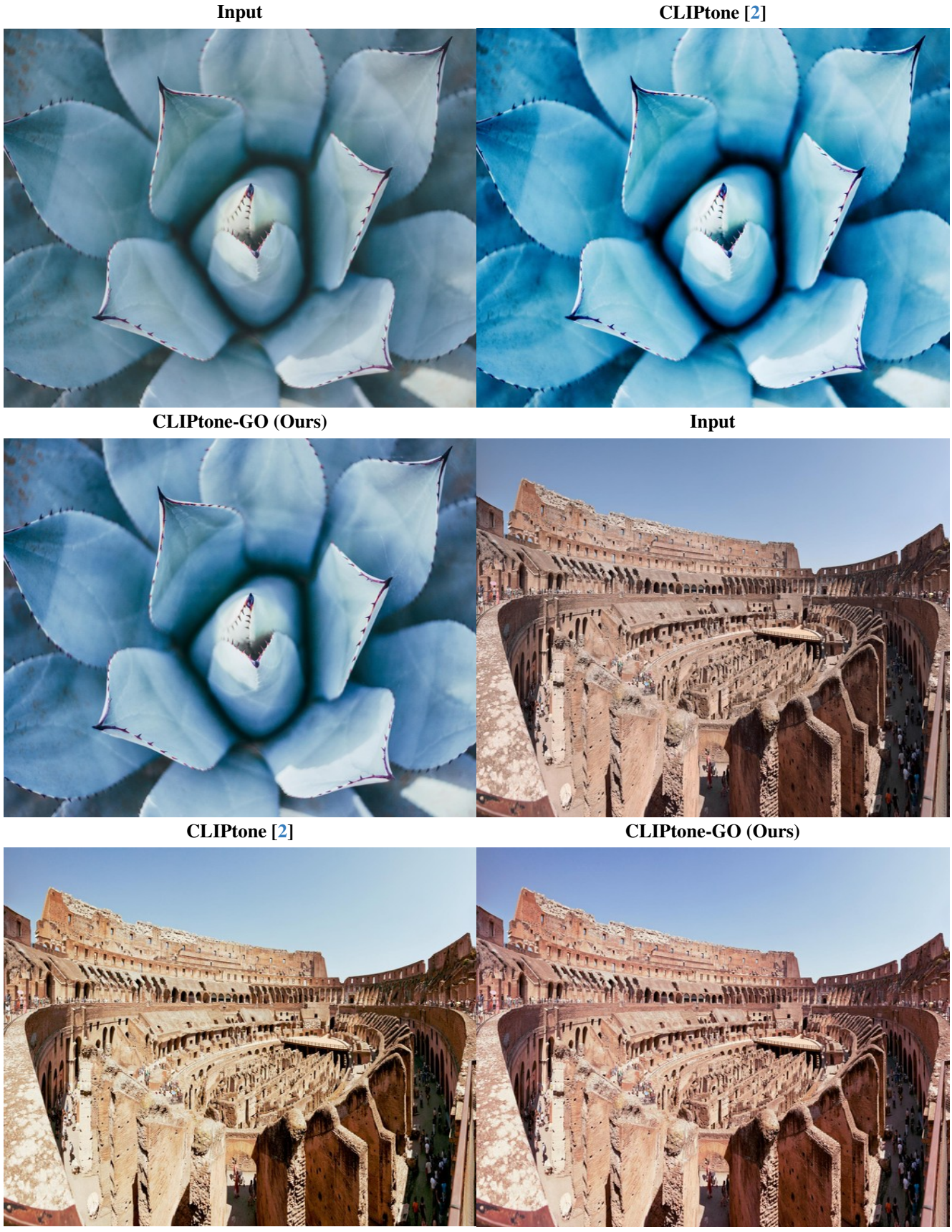


Figure 2. Extended View of Figure 1.

Figure 3. Extended View of Figure 1.

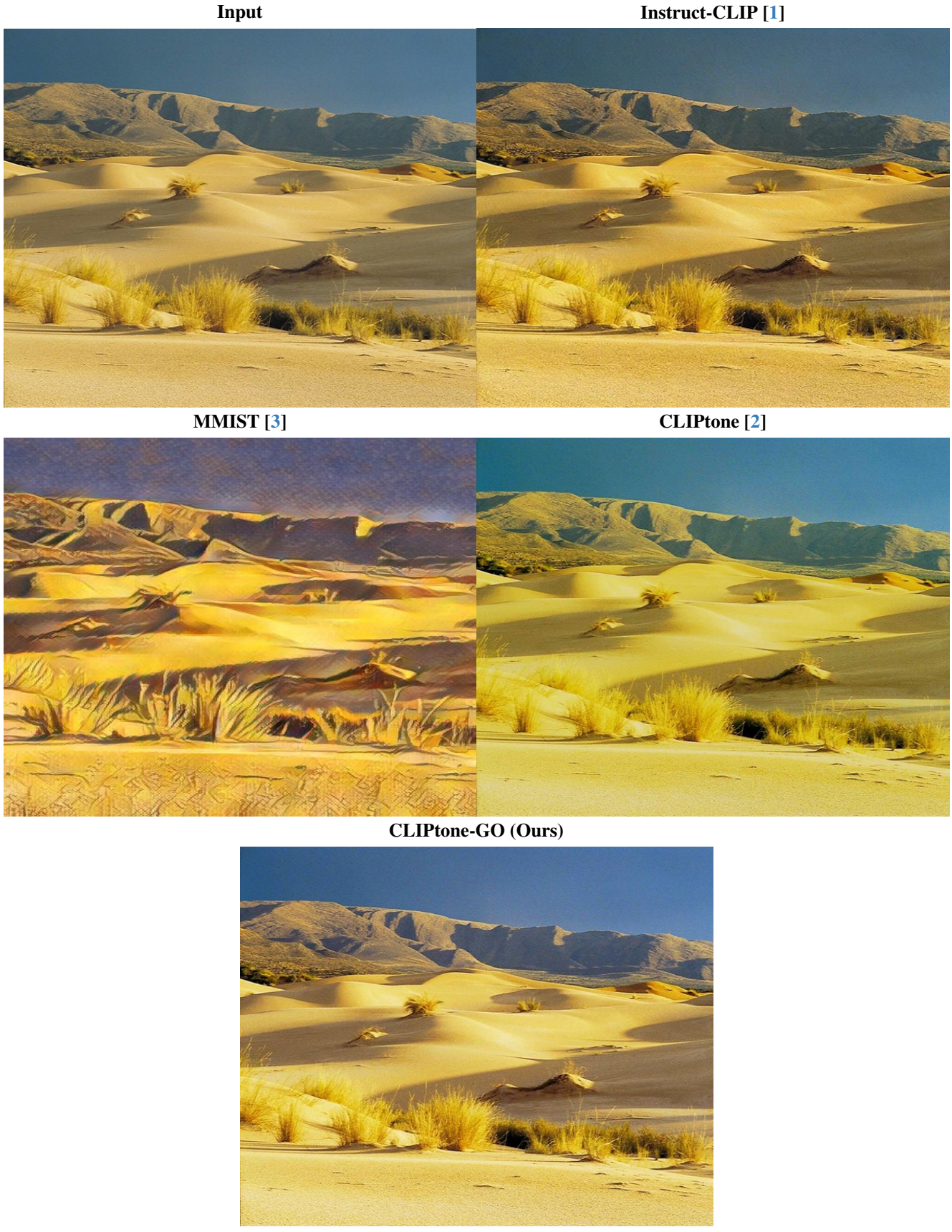


Figure 4. Extended View of Figure 3.

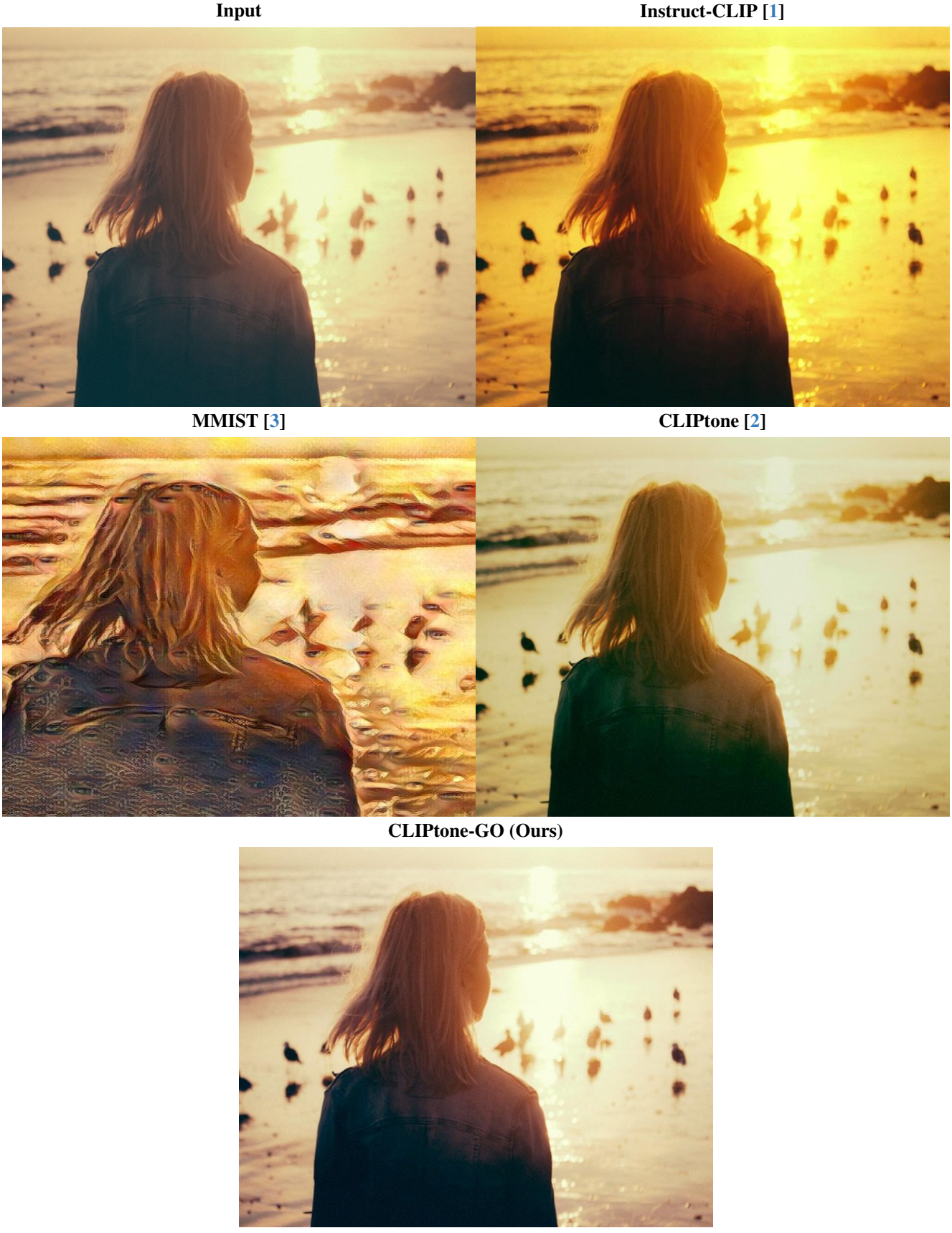


Figure 5. Extended View of Figure 3.

Input

Instruct-CLIP [1]

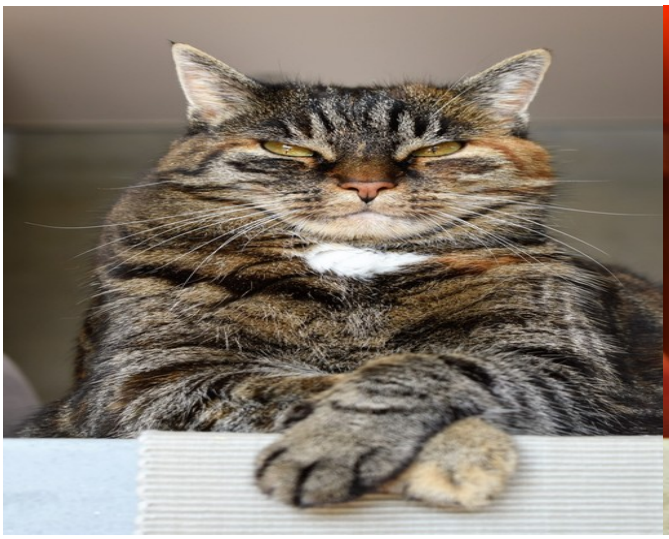
MMIST [3]

CLIPtone [2]

CLIPtone-GO (Ours)

Figure 6. Extended View of Figure 3.

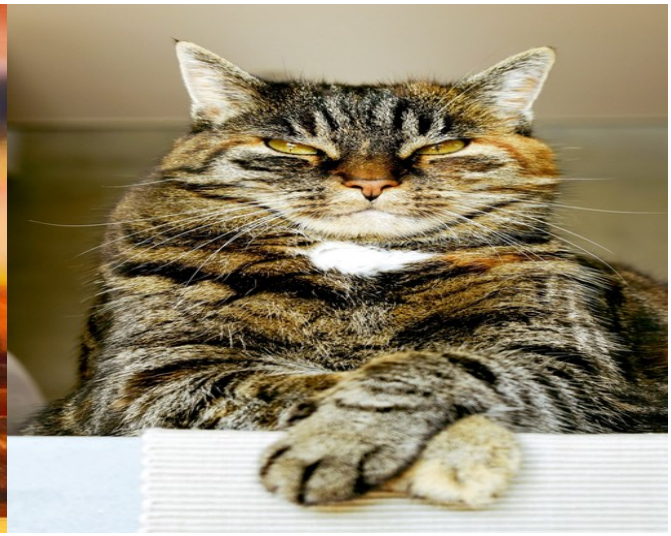
Input



Instruct-CLIP [1]

MMIST [3]

CLIPtone [2]



CLIPtone-GO (Ours)

Figure 7. Extended View of Figure 3.

Input **Instruct-CLIP [1]**

MMIST [3] **CLIPtone [2]**

CLIPtone-GO (Ours)

Figure 8. Extended View of Figure 3.

A. Adaptive CLIP Weight

This is the detailed proof of Property 2. Consider

$$f(\lambda) = \tfrac{1}{2}\left|g_0 + \lambda \tilde{g}_P\right|_2^2, \qquad \lambda \geq 0, \tag{1}$$

The unique minimizer is

$$\lambda^\star_{\text{CLIP}} = \max\left(0, -\frac{\langle g_0, \tilde{g}_P \rangle}{|\tilde{g}_P|_2^2}\right). \tag{2}$$

Proof (calculus). Expanding the square gives

$$f(\lambda) = \tfrac{1}{2}\Big(|g_0|_2^2 + 2\lambda\langle g_0, \tilde{g}_P\rangle + \lambda^2 |\tilde{g}_P|_2^2\Big). \tag{3}$$

Thus f is a convex quadratic in λ with

$$f'(\lambda) = \langle g_0, \tilde{g}_P\rangle + \lambda\|\tilde{g}_P\|_2^2, \quad f''(\lambda) = \|\tilde{g}_P\|_2^2 \geq 0. \tag{4}$$

The unconstrained stationary point solves $f'(\lambda) = 0$, yielding

$$\bar{\lambda} = -\frac{\langle g_0, \tilde{g}_P\rangle}{|\tilde{g}_P|_2^2}. \tag{5}$$

Imposing the constraint $\lambda \geq 0$ amounts to projecting the unconstrained minimizer onto $[0, \infty)$:

$$\lambda^\star_{\text{CLIP}} = \max(0, \bar{\lambda}) = \max\left(0, -\frac{\langle g_0, \tilde{g}_P\rangle}{\|\tilde{g}_P\|_2^2}\right), \tag{6}$$

which proves (2). In the degenerate case $\tilde{g}_P = 0$, $f(\lambda) \equiv \frac{1}{2}\|g_0\|_2^2$ is constant and any $\lambda \geq 0$ minimizes f; we take $\lambda^\star_{\text{CLIP}} = 0$ by convention.

B. Visualization of Geometry-Aware Training

It is visualization of the real values of $\lambda^\star_{\text{LPIPS}}$ and α_{inst}.

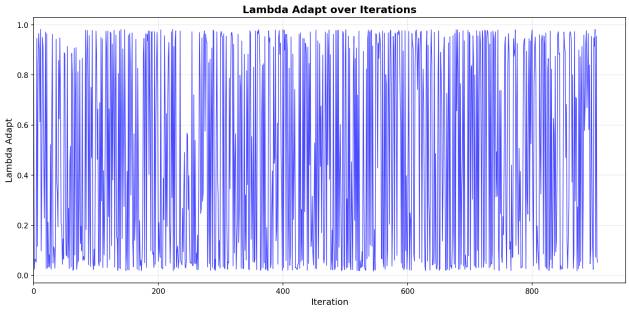


Figure 9. Lambda Adapt Plot

As shown in the $\lambda^\star_{\text{LPIPS}}$ plot (blue), the optimal strength for the perceptual correction is not static. It oscillates rapidly between negligible influence (< 0.1) and strong enforcement (> 0.9). This confirms that the agreement between CLIP and LPIPS gradients changes instantaneously.

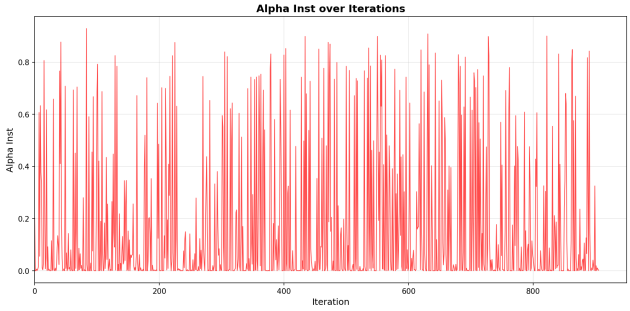


Figure 10. Lambda Inst Plot

The α_{inst} plot (red) illustrates our model's trust mechanism. The sparsity of high values indicates that the model frequently defaults to the conservative Adaptive CLIP weight to resolve conflicts and only to switch to the aggressive LPIPS weight (high α) when the gradient geometry strictly permits it.

References

[1] Sherry X. Chen, Misha Sra, and Pradeep Sen. Instruct-clip: Improving instruction-guided image editing with automated data refinement using contrastive learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28513–28522, 2025. 4, 5, 6, 7, 8

[2] Hyeongmin Lee, Kyoungkook Kang, Jungseul Ok, and Sunghyun Cho. Cliptone: Unsupervised learning for text-based image tone adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2942–2951, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[3] Hanyu Wang, Pengxiang Wu, Kevin Dela Rosa, Chen Wang, and Abhinav Shrivastava. Multimodality-guided image style transfer using cross-modal gan inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4976–4985, 2024. 4, 5, 6, 7, 8