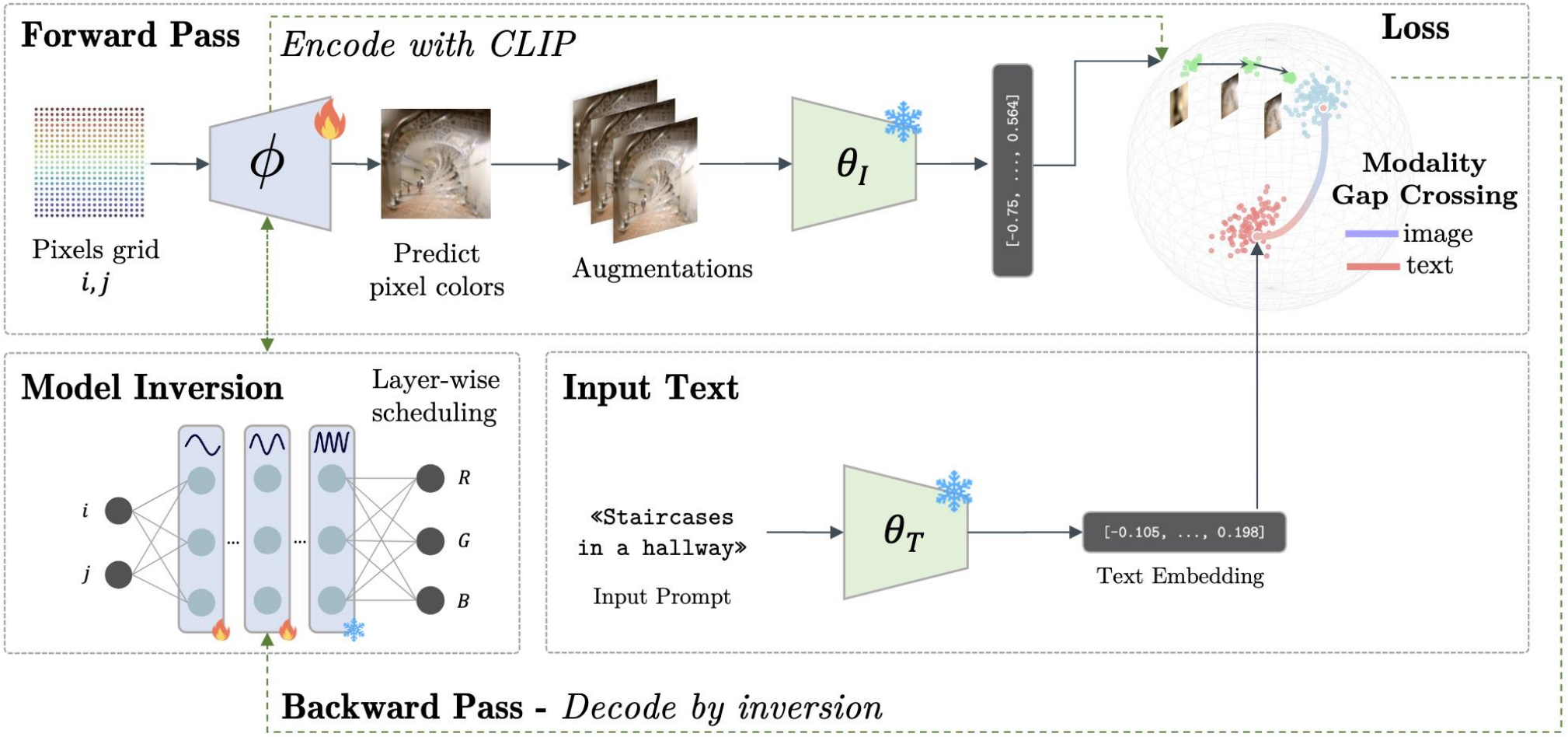
Implicit Inversion turns CLIP into a Decoder

Antonio D'Orazio[1], Maria Rosaria Briglia[1], Donato Crisostomi[2,1], Dario Loi[1], Emanuele Rodolà[2,3], Iacopo Masi[1]

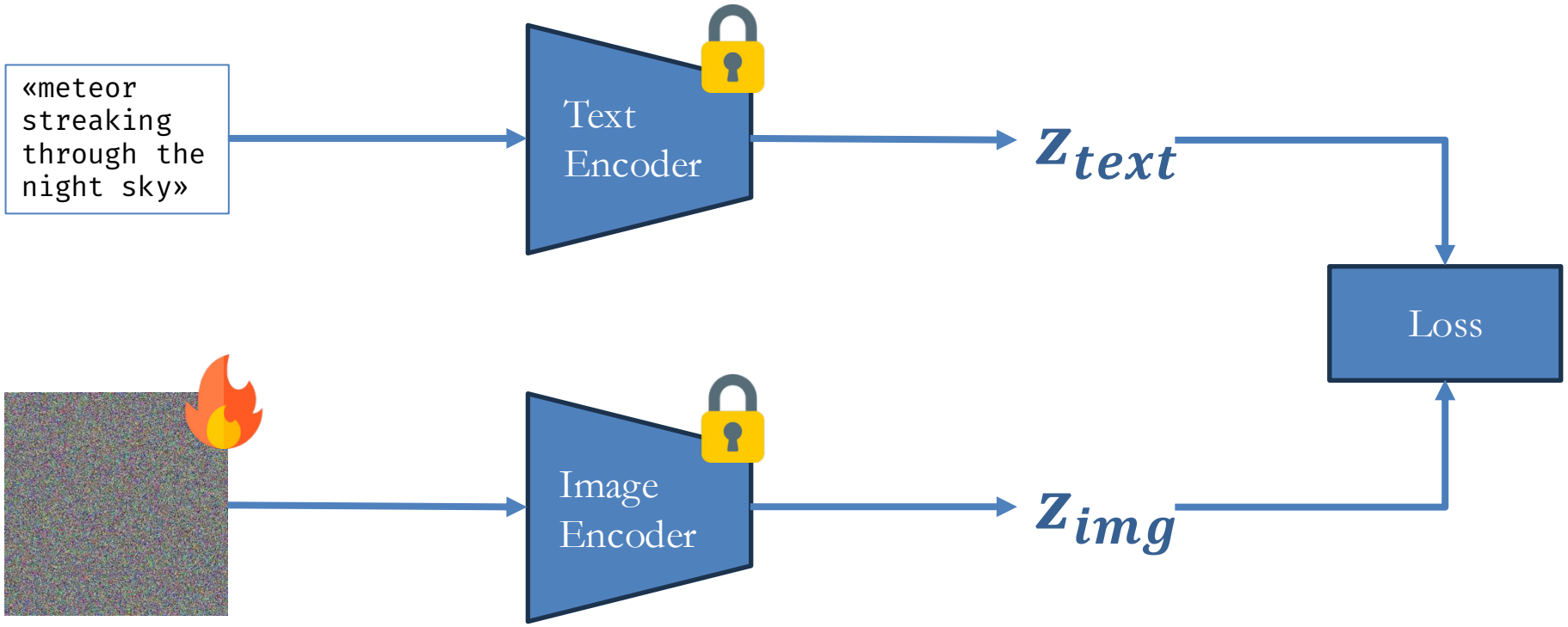
Sapienza, University of Rome, Italy
Department of Computer Science

[1]OmnAI, [2]GLADIA, [3]Paradigma

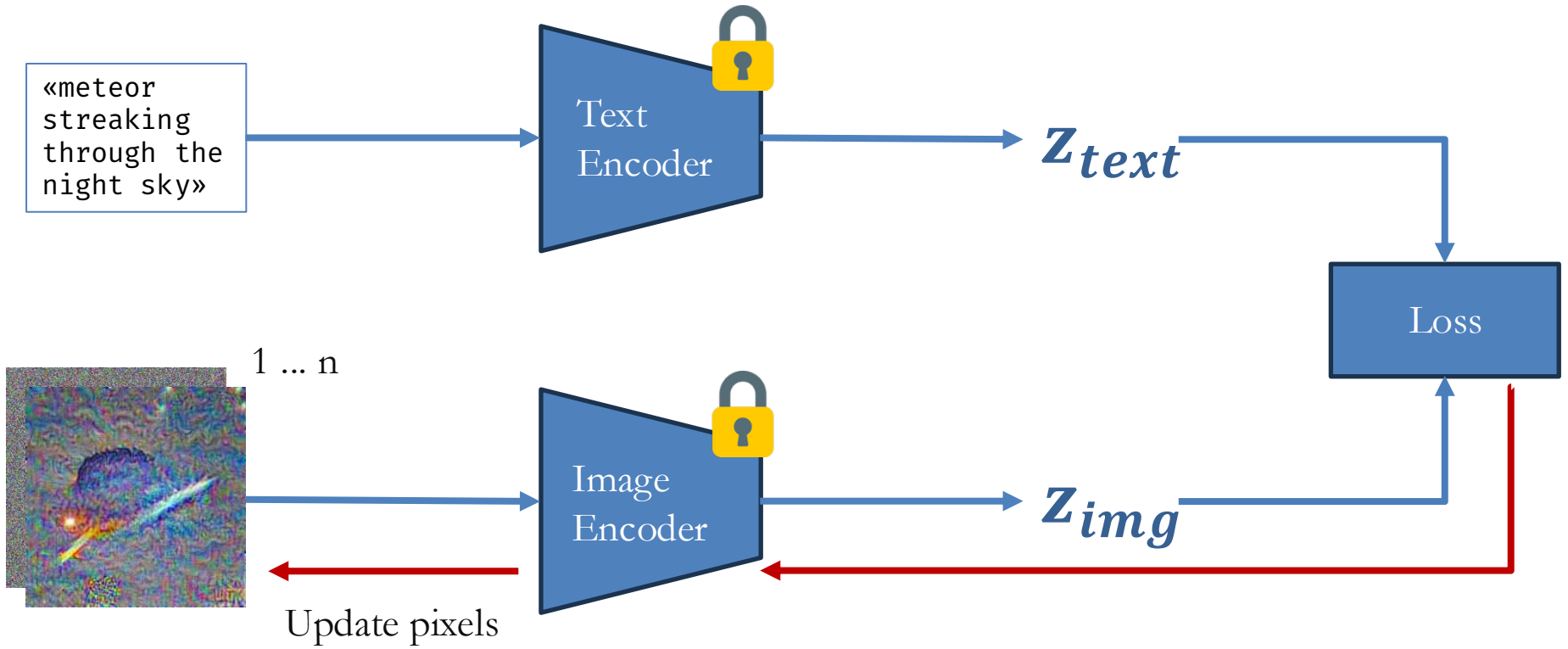
Method


Forward Pass
Encode with CLIP
Loss
Pixels grid
i, j
ϕ
Predict
pixel colors
Augmentations
θ_I
[-0.75, ..., 0.564]
Modality
Gap Crossing
image
text
Model Inversion
Layer-wise
scheduling
i
j
R
G
B
Input Text
«Staircases
in a hallway»
Input Prompt
θ_T
[-0.105, ..., 0.198]
Text Embedding
Backward Pass - Decode by inversion

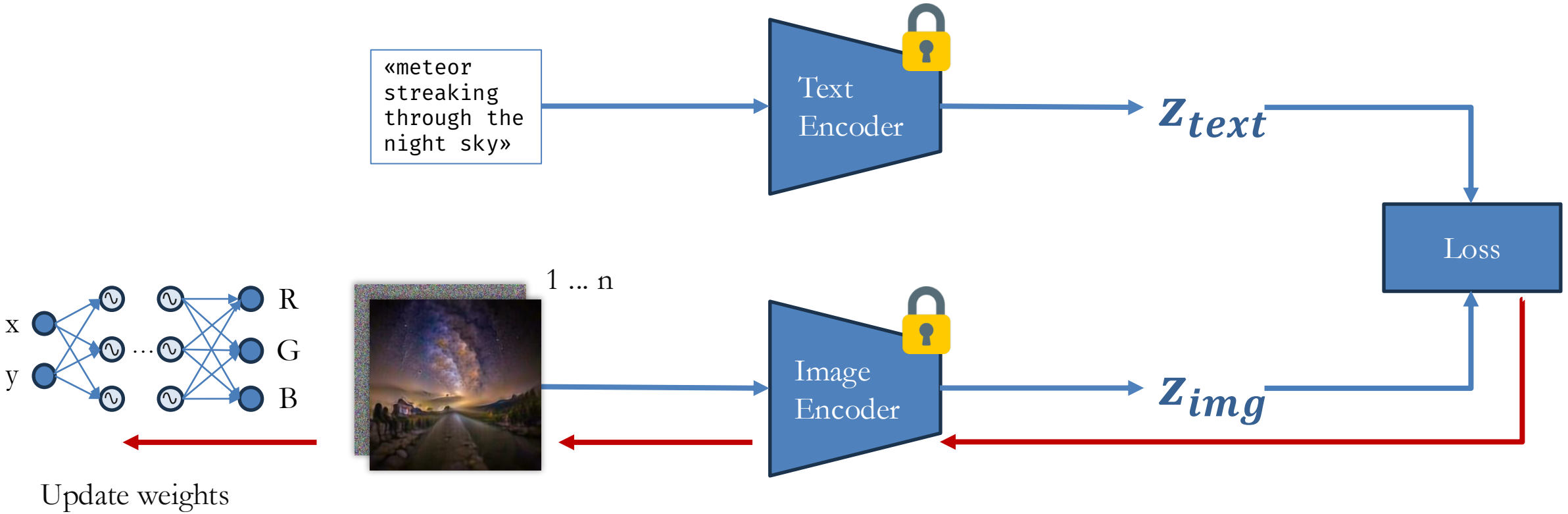
Encoder Only Image Synthesis



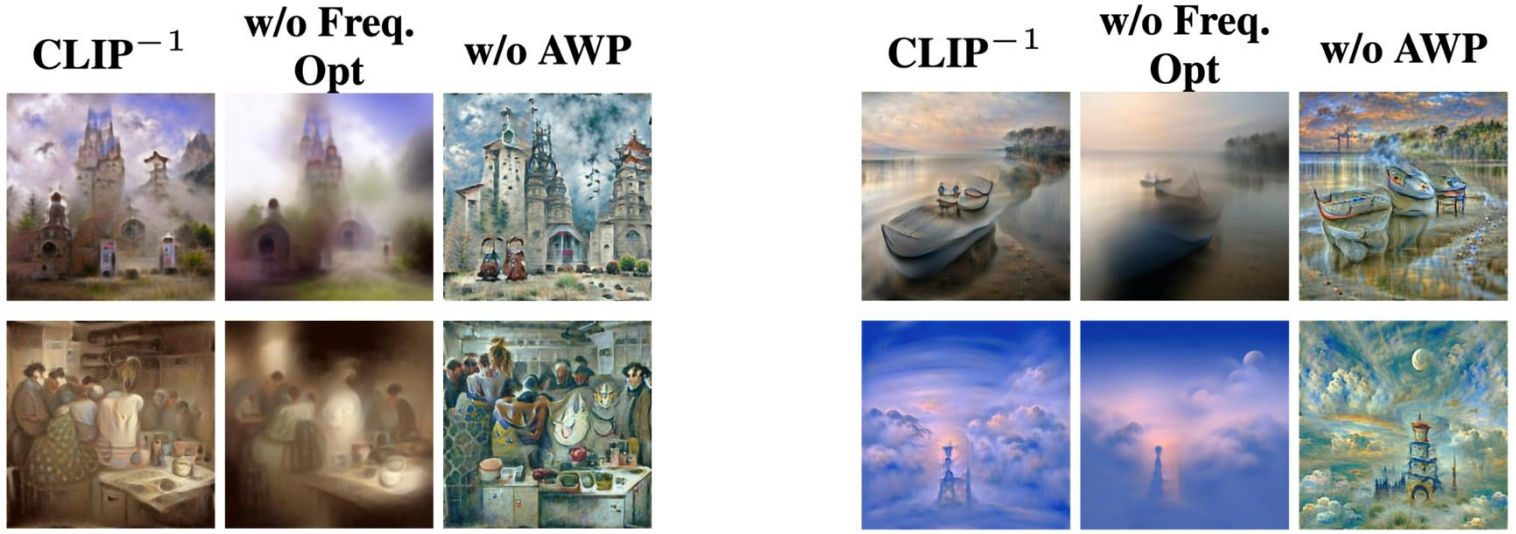
Encoder Only Image Synthesis: Inversion



Encoder Only Image Synthesis: Implicit inversion

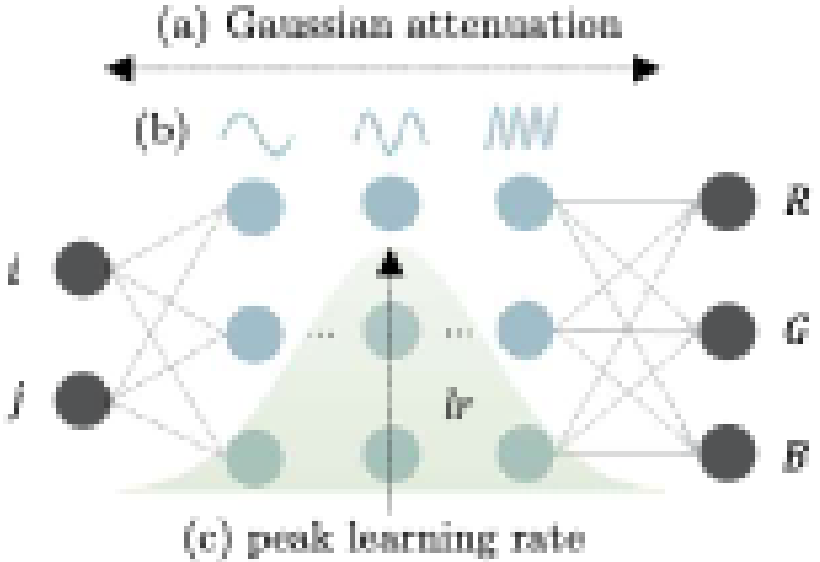


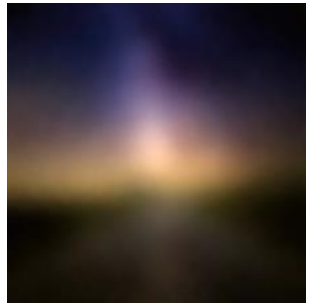
Stable Init with Adversarial Weights Perturbation



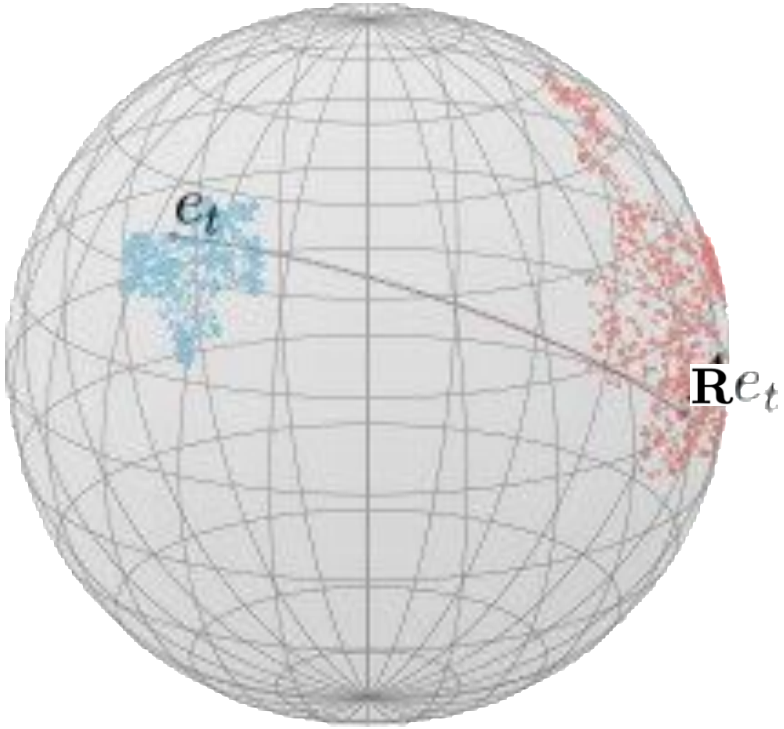
$$\min_{\phi} \max_{\Delta\phi \in \Omega} \mathcal{L}\left(f_{\phi+\Delta\phi}, \mathbf{blur}(\mathbf{x})\right)$$

Optimize by frequency



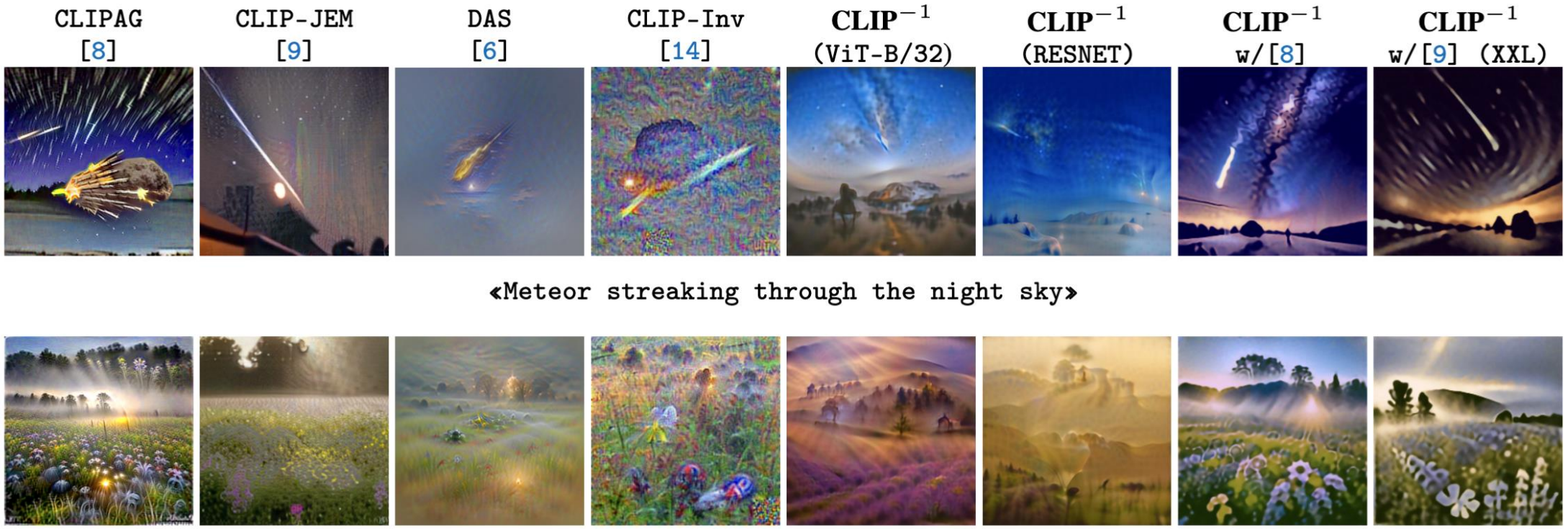


Modality Gap Crossing

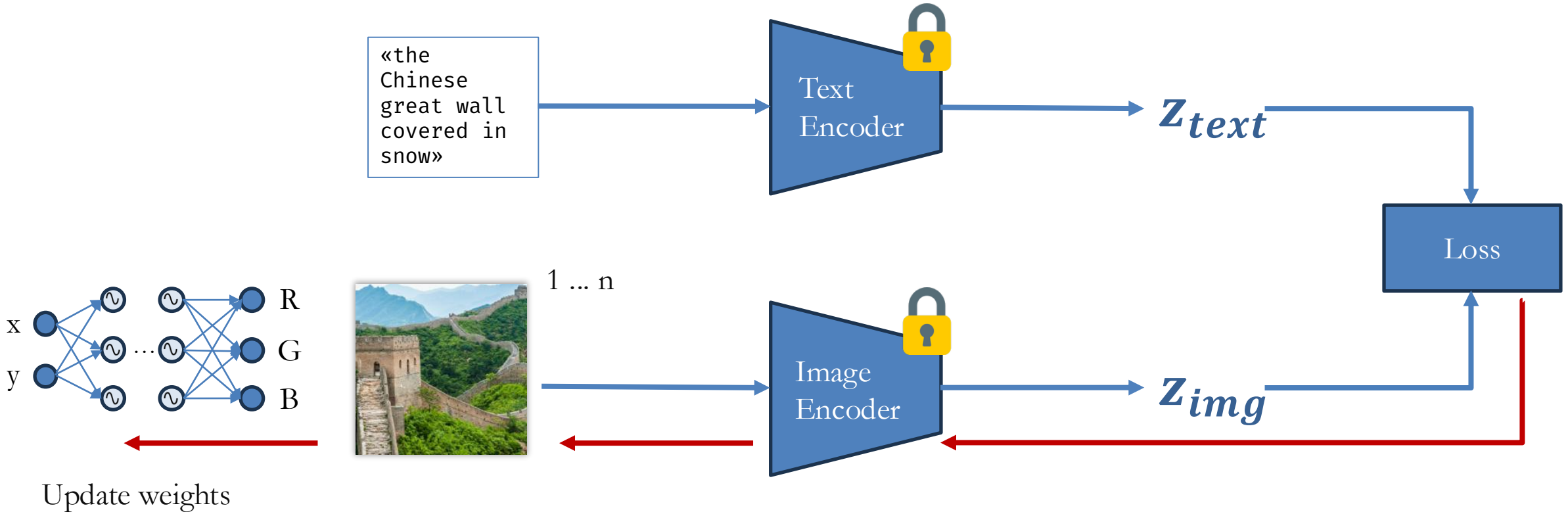


$$\min_{\mathbf{R}} ||\mathbf{R}\mathbf{E}_T - \mathbf{E}_I||_F \qquad \text{s.t. } \mathbf{R}^\top \mathbf{R} = \mathbf{I}$$

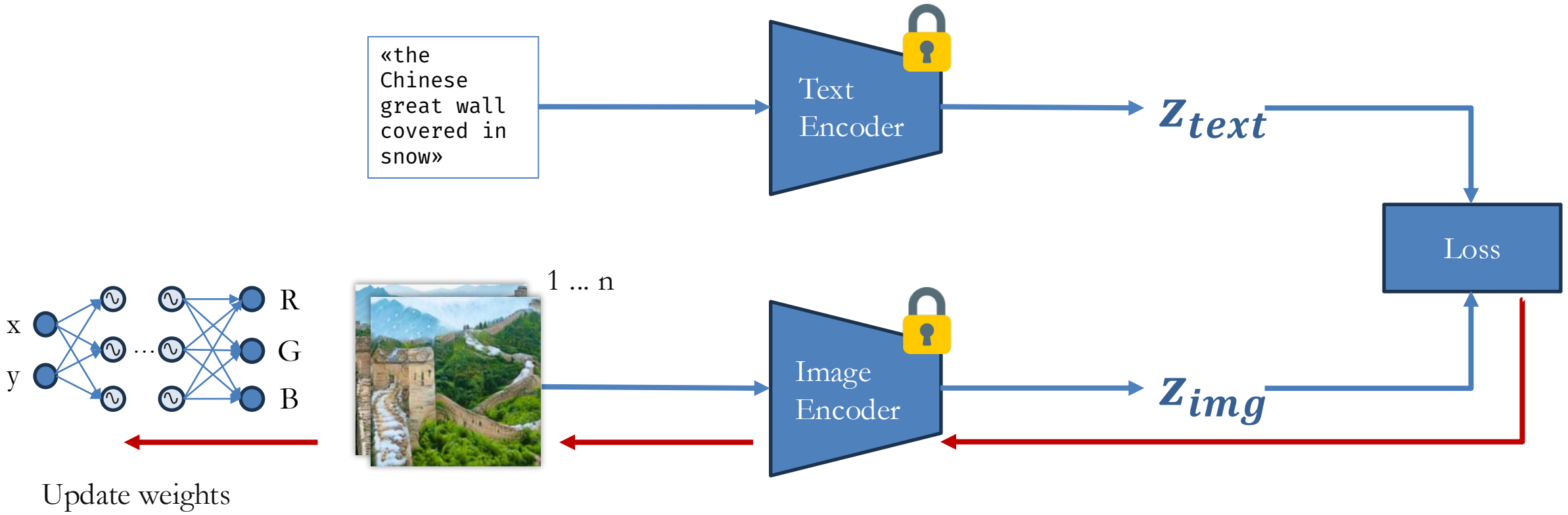
Quantitative Comparison



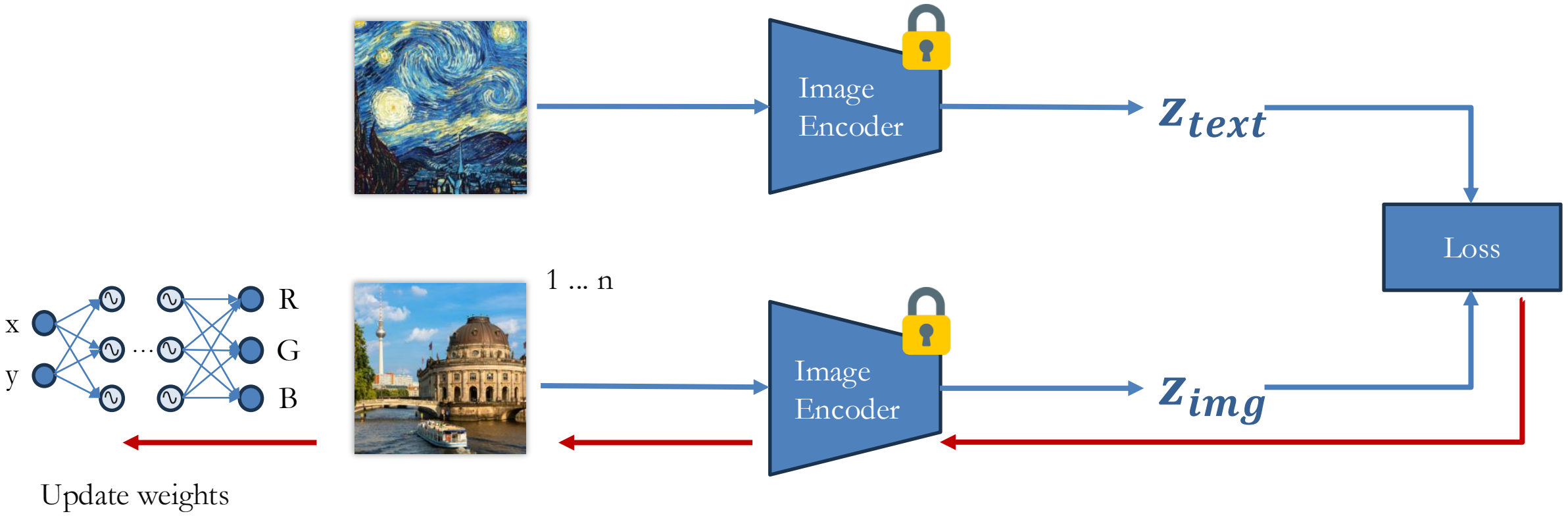
Downstream Tasks: Controlled modification



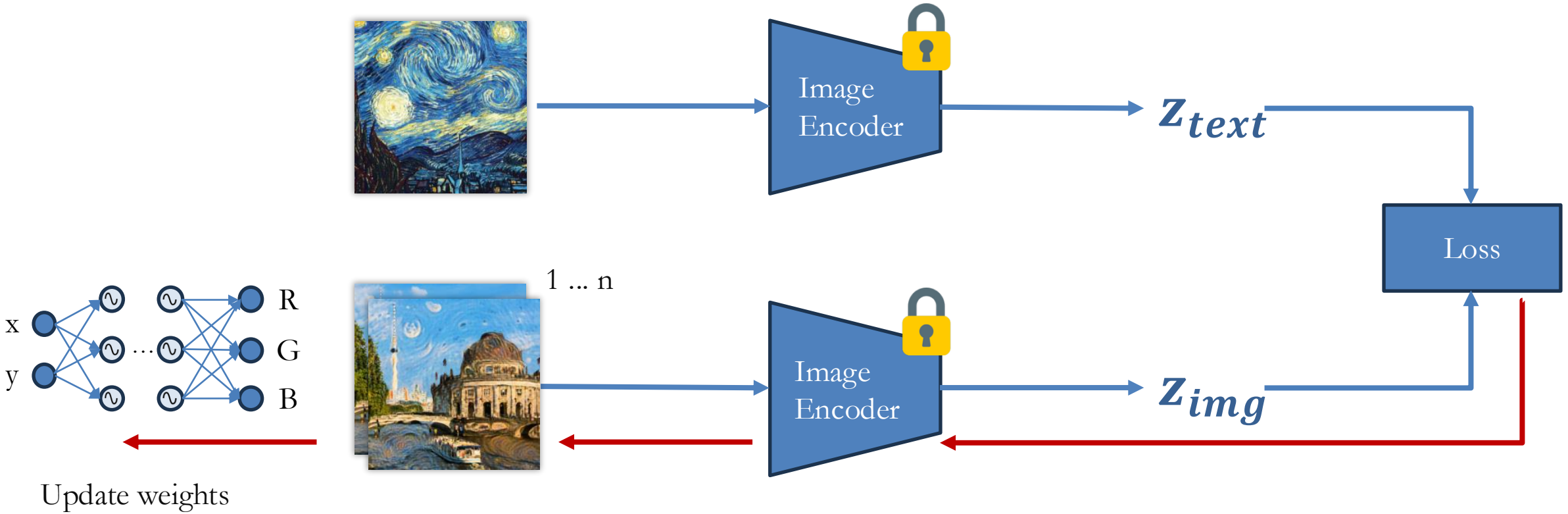
Downstream Tasks: Controlled modification



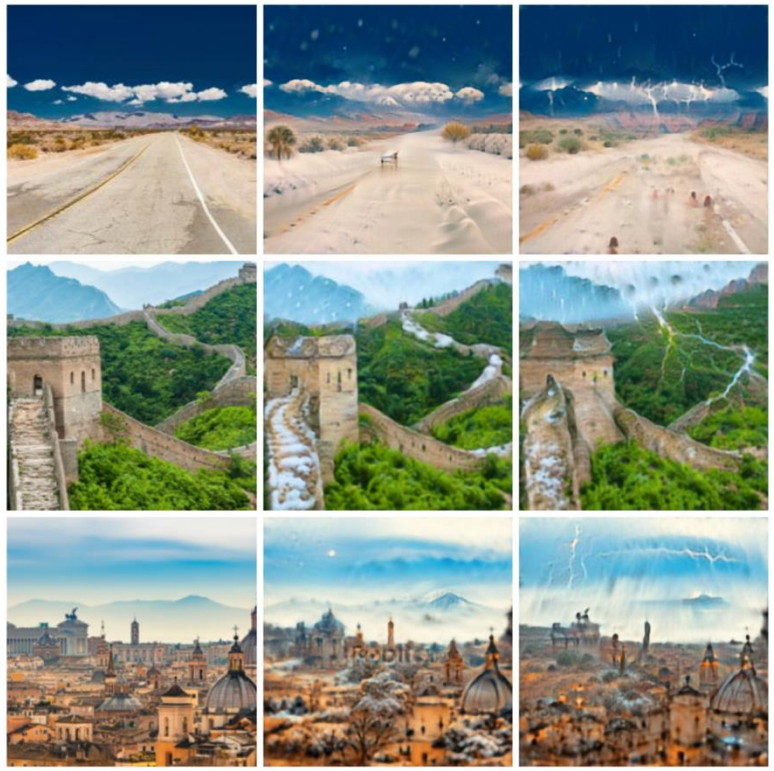
Downstream Tasks: Style Transfer



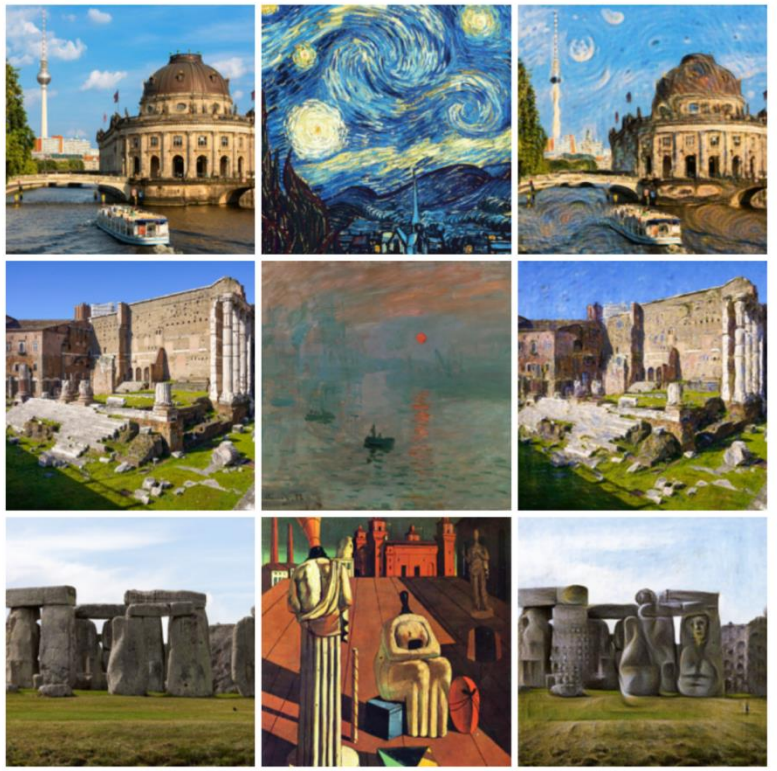
Downstream Tasks: Style Transfer



Results

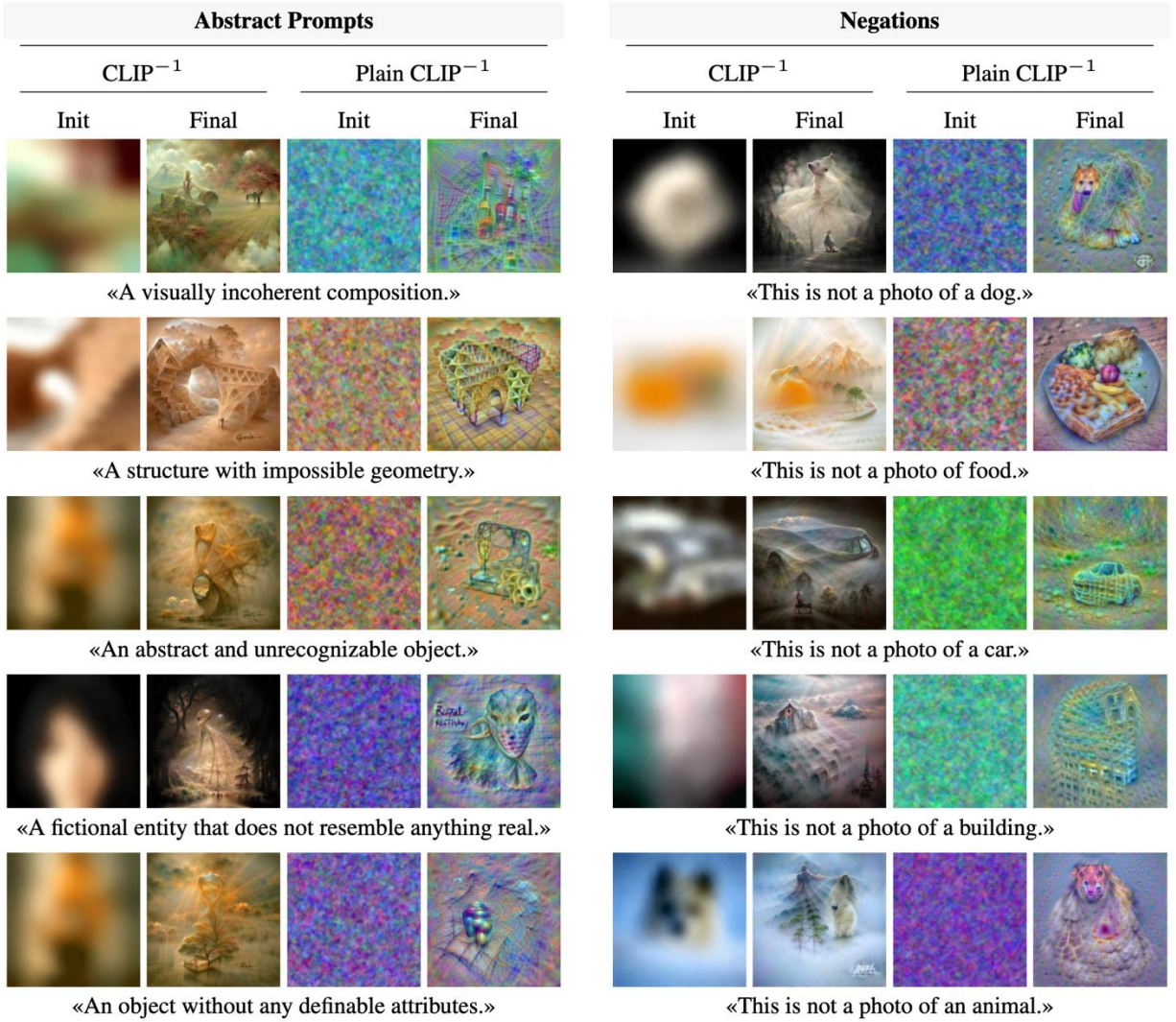


Original Prompt 1 Prompt 2



Original Reference Result

Interpretability of CLIP's embedding space



Thank you!