

Implicit Inversion turns CLIP into a Decoder



Mt Pinatubo eruption

Tibetan monastery on a cliff

Prague castle

Machu Picchu shrouded in mist

Grand Canyon at sunset

Photo of Uluru



A group of people sitting around a table

Kept retro 1969s looking living room set

A pastry store with cupcakes on display

Some oranges are stacked up in a bowl

A man standing near a blue and brown bed

A kitchen with cabinets and stuff on the table

Antonio D'Orazio¹, Maria Rosaria Briglia¹, Donato Crisostomi^{2,1},
Dario Loi¹, Emanuele Rodolà^{2,3}, Iacopo Masi¹



Sapienza, University of Rome, Italy
Department of Computer Science



¹OmnAI,

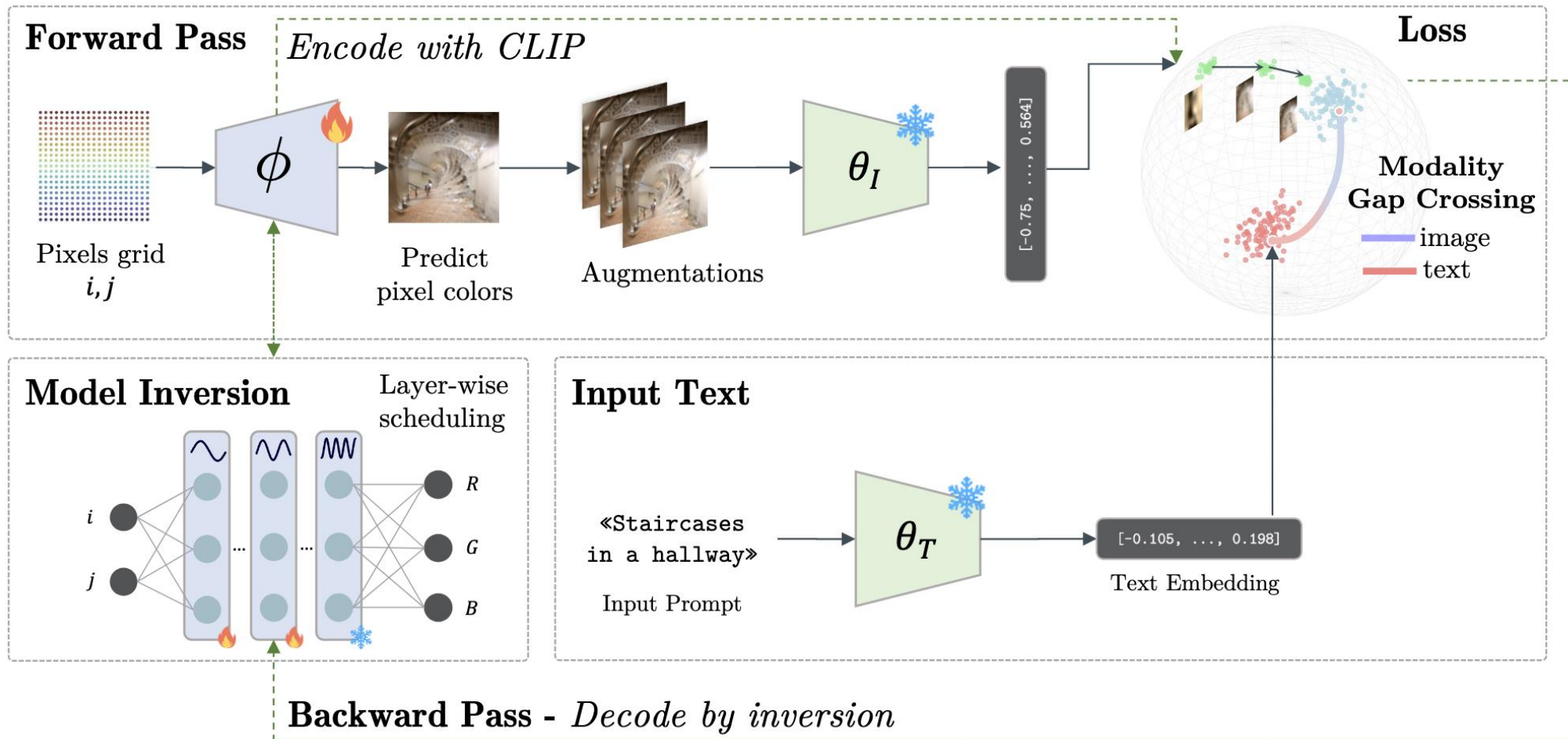


²GLADIA,

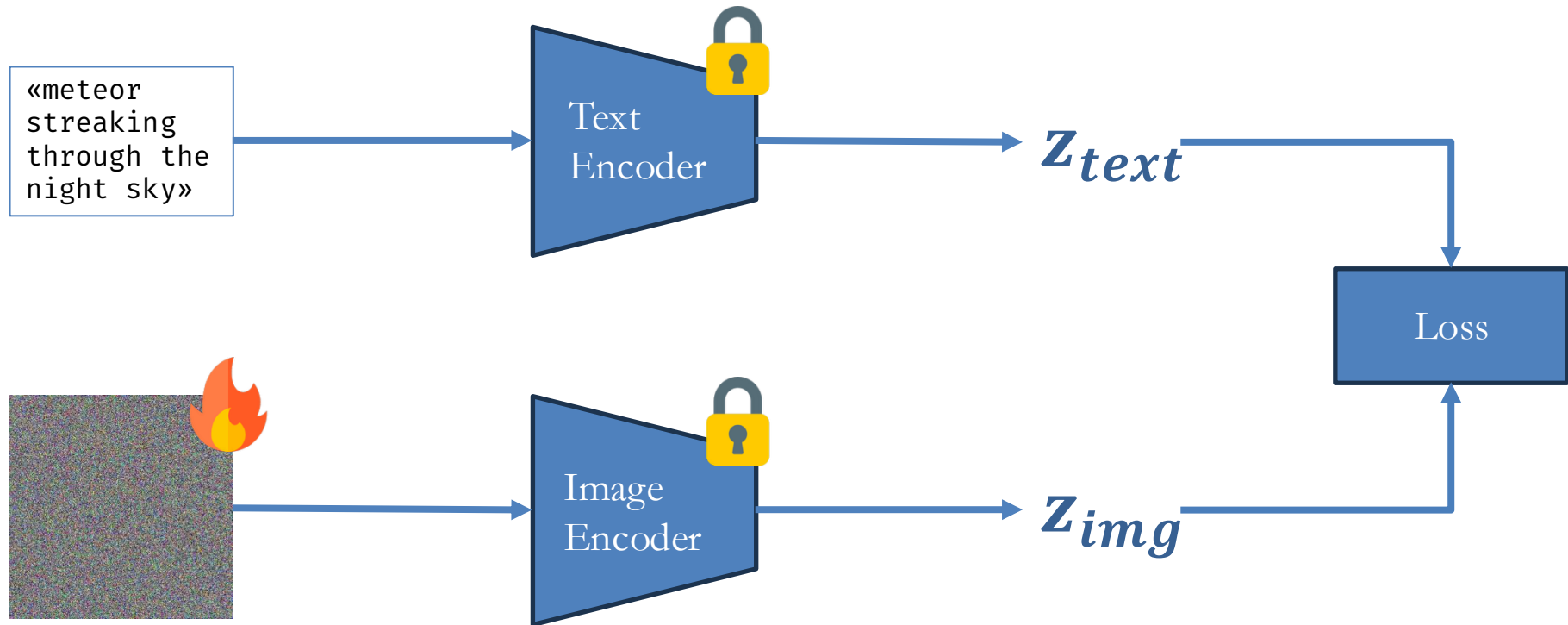


³Paradigma

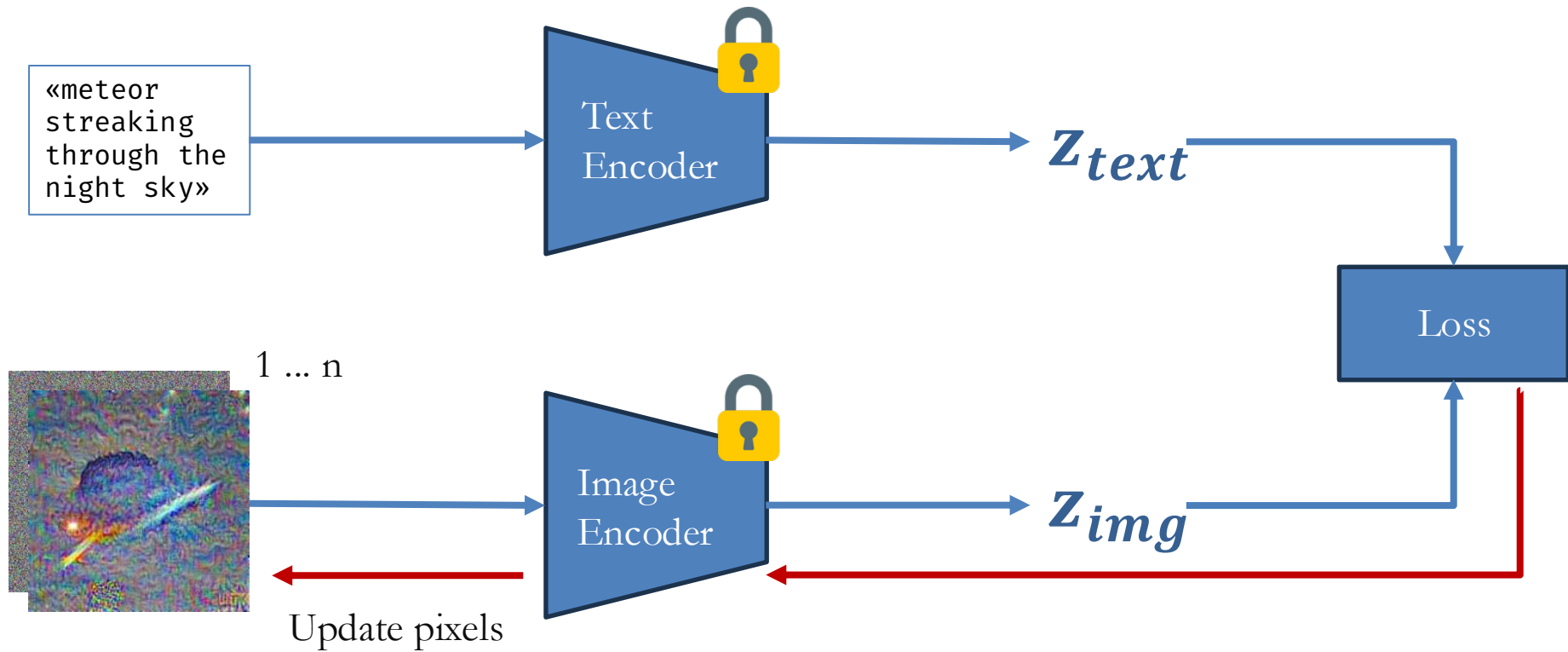
Method



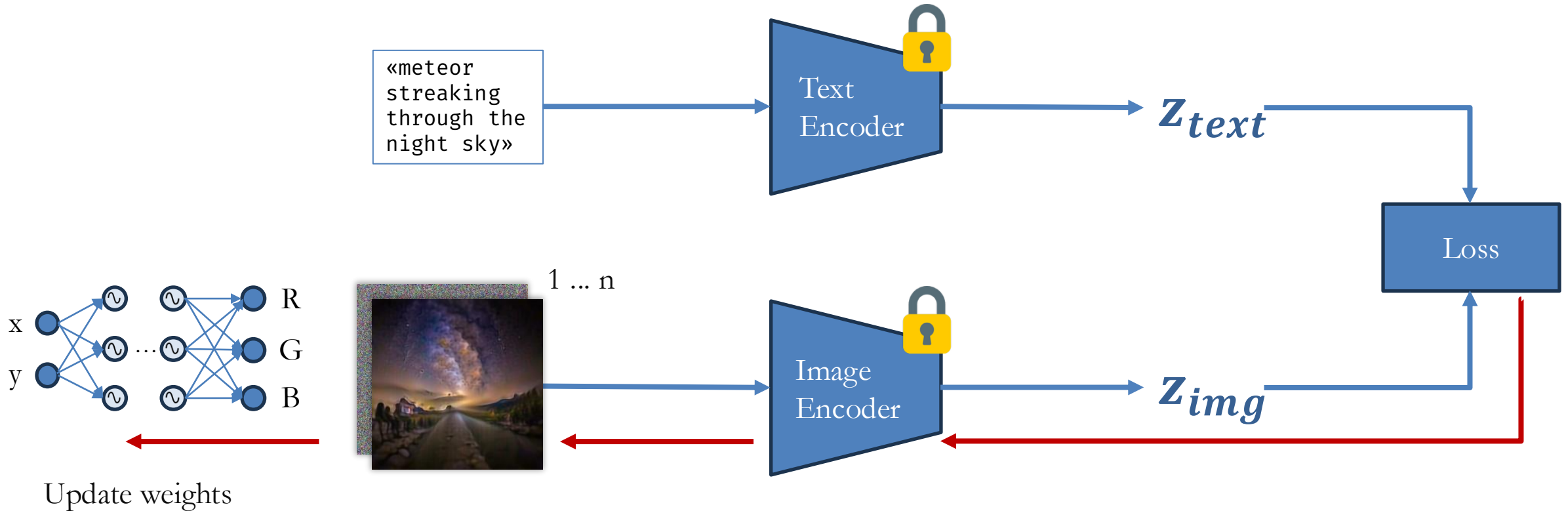
Encoder Only Image Synthesis



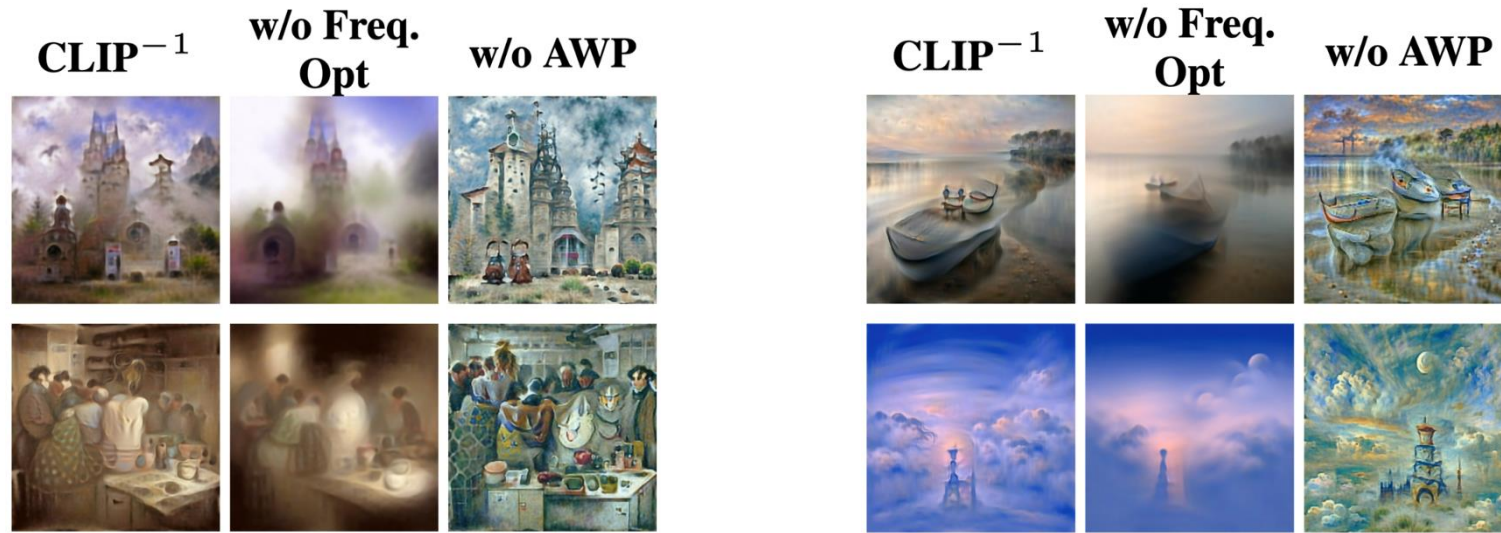
Encoder Only Image Synthesis: Inversion



Encoder Only Image Synthesis: Implicit inversion



Stable Init with Adversarial Weights Perturbation

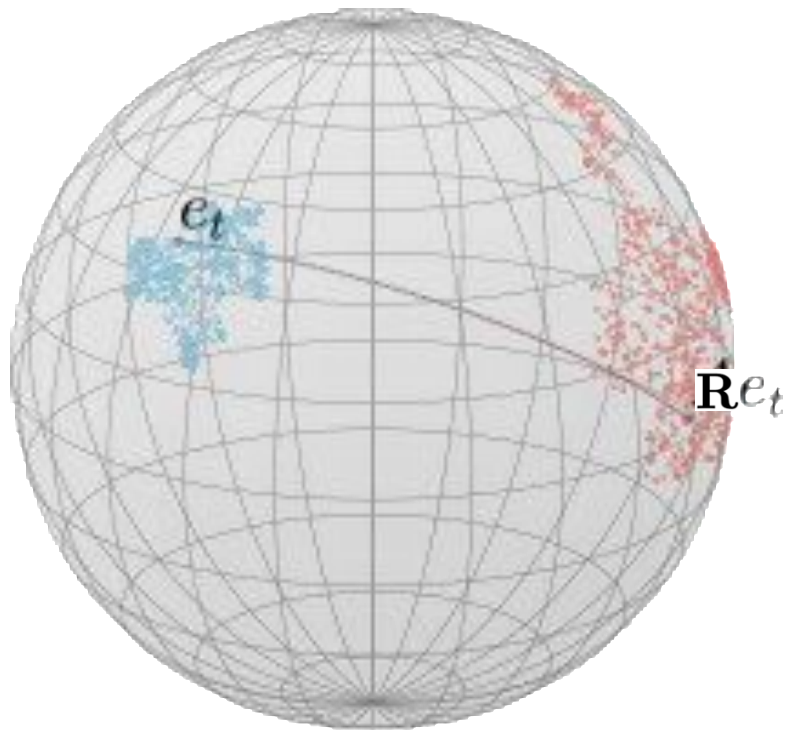


$$\min_{\phi} \max_{\Delta\phi \in \Omega} \mathcal{L}(f_{\phi + \Delta\phi}, \text{blur}(\mathbf{x}))$$

Optimize by frequency



Modality Gap Crossing



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

Quantitative Comparison

CLIPAG
[8]



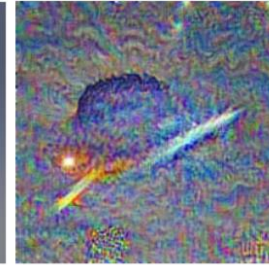
CLIP-JEM
[9]



DAS
[6]



CLIP-Inv
[14]



CLIP⁻¹
(ViT-B/32)



CLIP⁻¹
(RESNET)



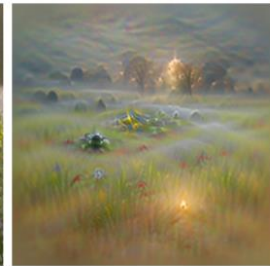
CLIP⁻¹
w/[8]



CLIP⁻¹
w/[9] (XXL)

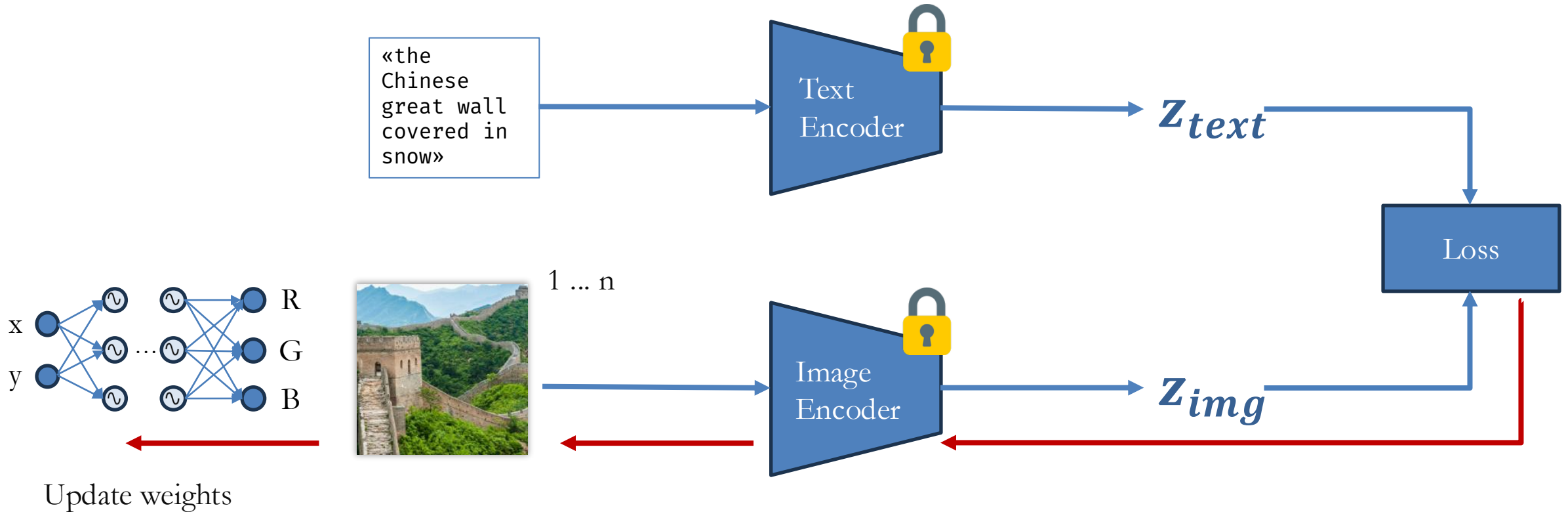


«Meteor streaking through the night sky»

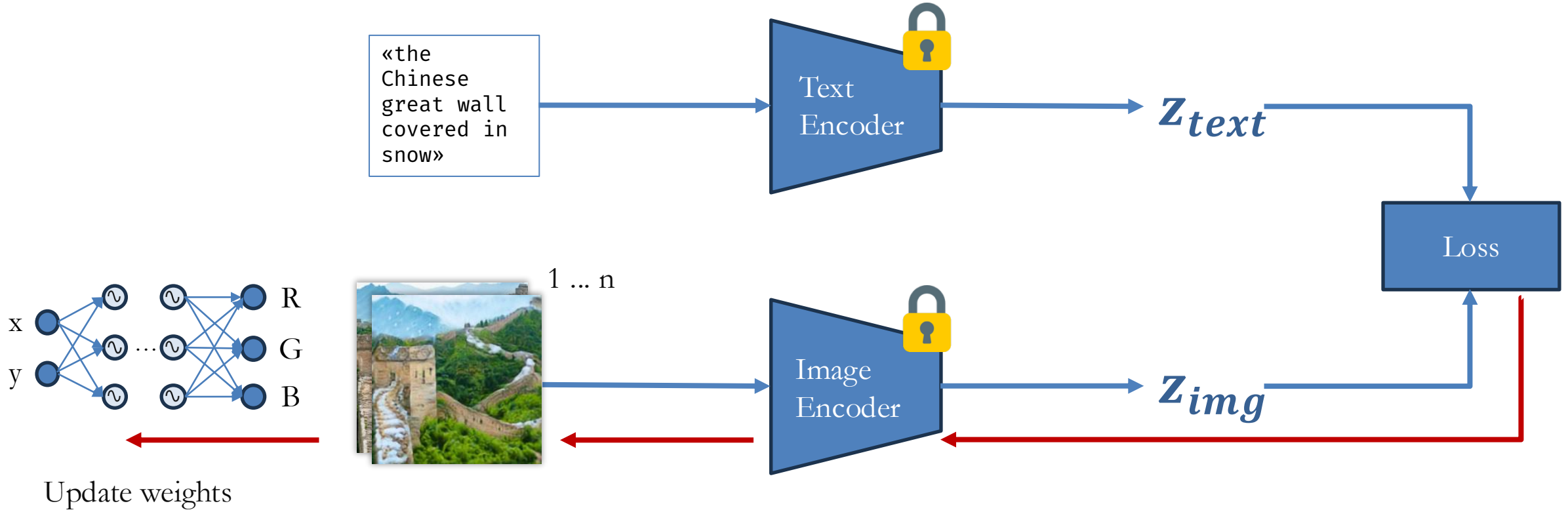


«A mist-covered field at daybreak with wildflowers glistening in early rays.»

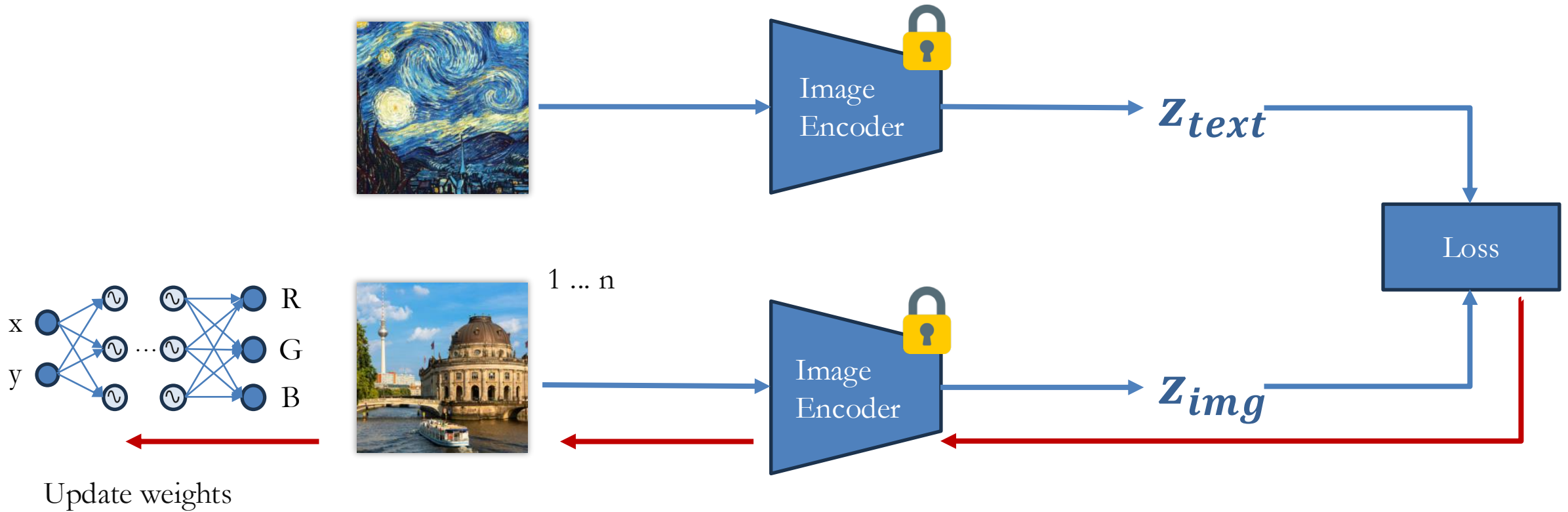
Downstream Tasks: Controlled modification



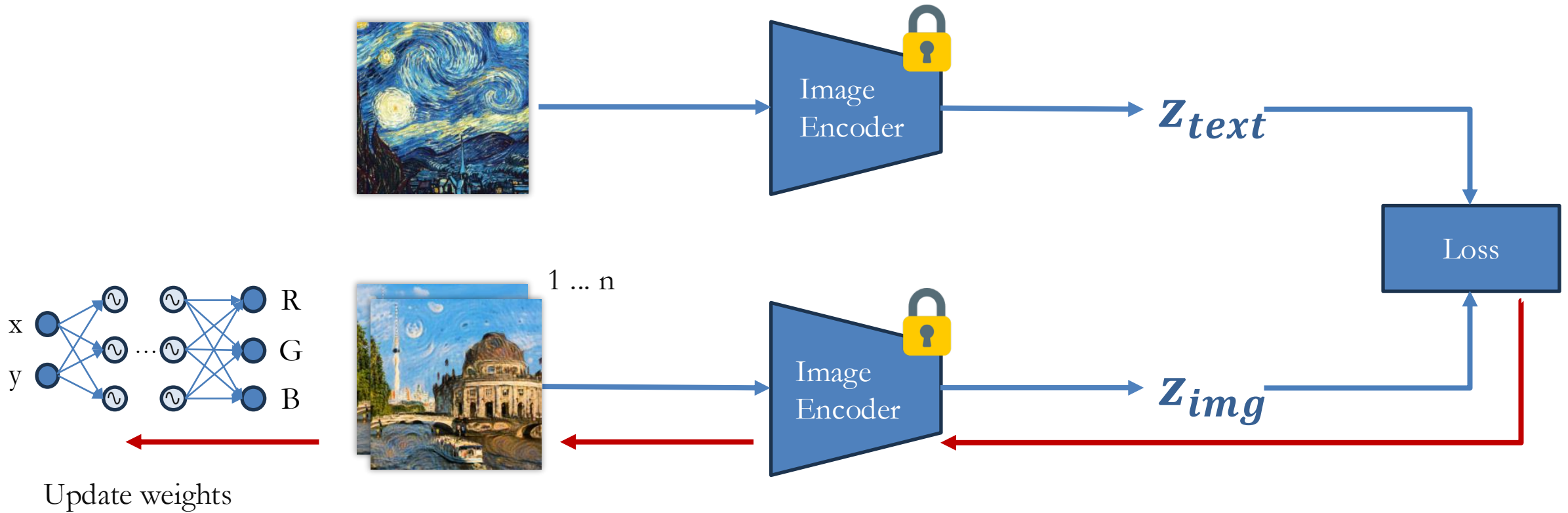
Downstream Tasks: Controlled modification



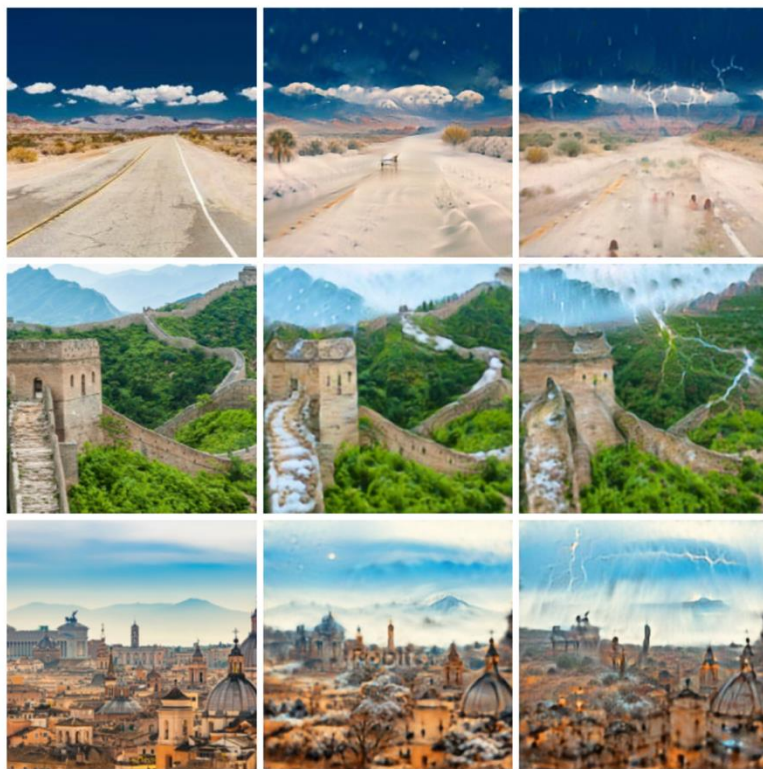
Downstream Tasks: Style Transfer



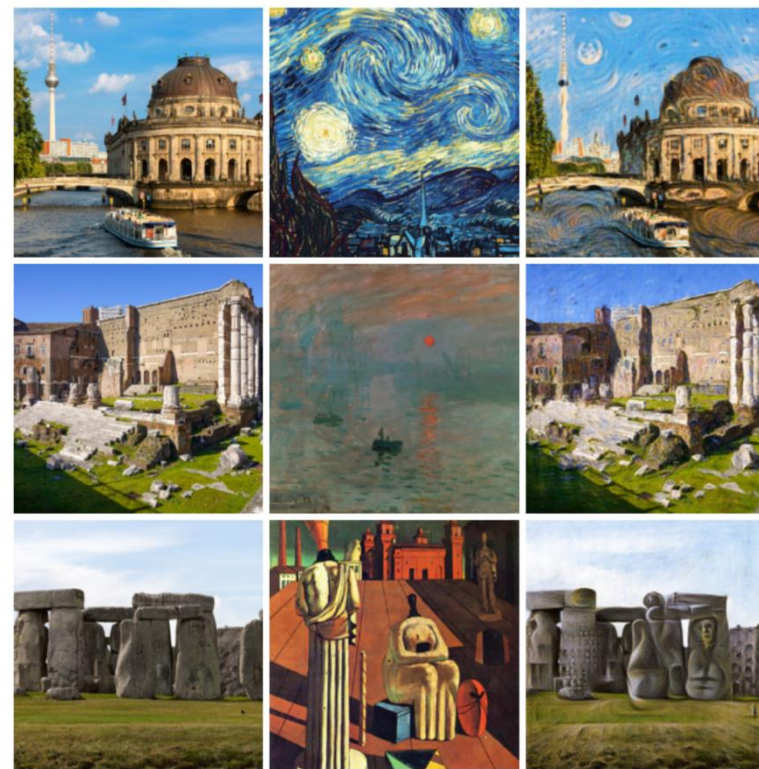
Downstream Tasks: Style Transfer



Results

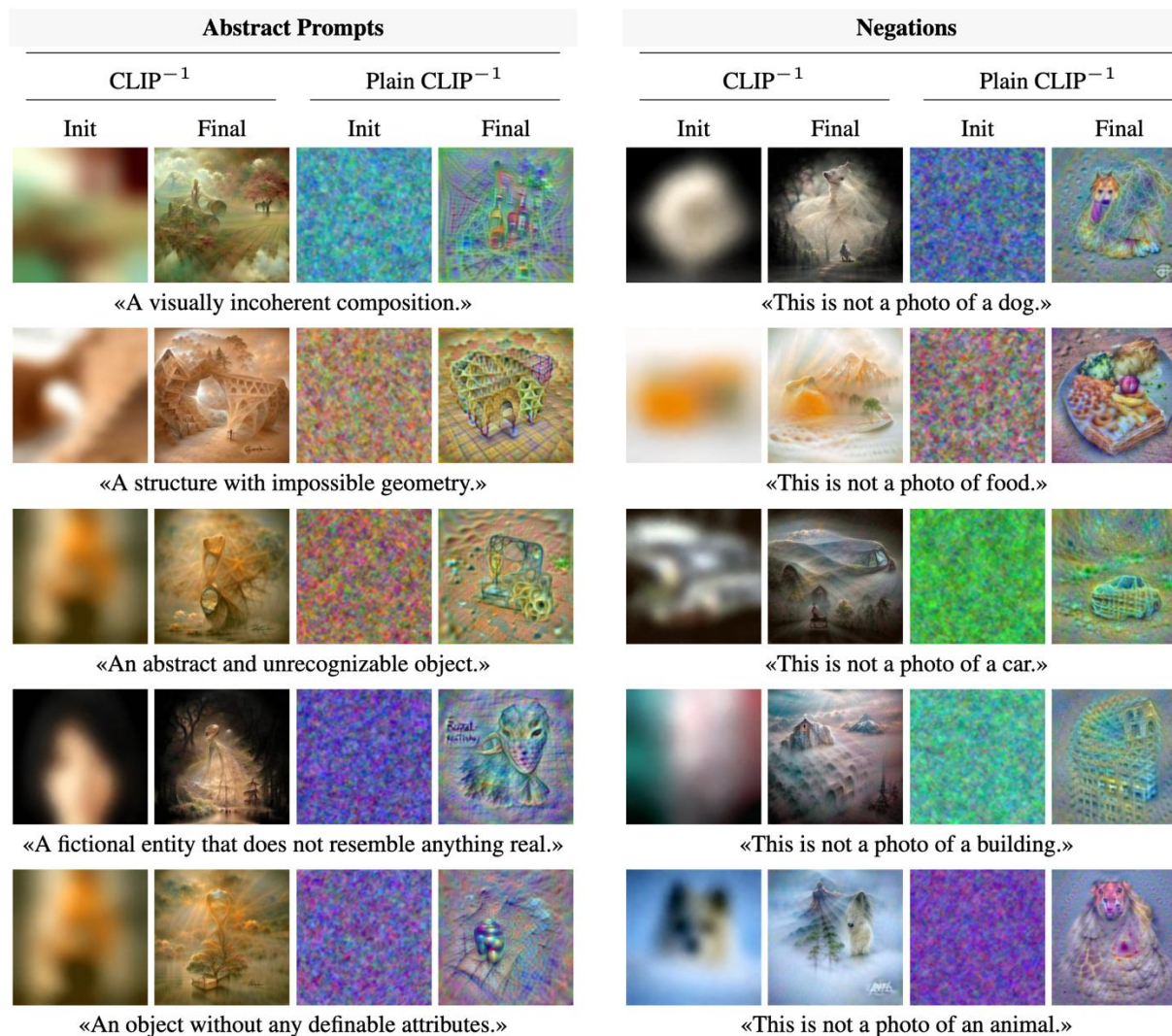


Original Prompt 1 Prompt 2



Original Reference Result

Interpretability of CLIP's embedding space



Thank you!