A Multi-Reference Style and Multi-Modal Context-Awareness Zero-Shot Style Alignment for Image Generation

***Alessio Borgi**, Luca Maiano, Irene Amerini*

Sapienza University of Rome, Italy

ABSTRACT

We propose a novel framework:

- **Zero-Shot Style Alignment** in **Image generation**.
- Incorporates **Multi-Modal Context-Awareness.**
- Incorporates **Multi-Reference Style Alignment**.

Unlike traditional methods, which rely solely on text input and require fine-tuning for style consistency, our approach:

- I**ntegrates diverse content Sources (images, audio, weather data,** and **music),** using models such as **BLIP-1**, **Whisper**, and **CLAP** to generate Multi-Modal textual description Embeddings.
- **Multiple Reference Images** and advanced **Blending Techniques** like **Linear Weighted Blending** and **Spherical Interpolation**.

ATTENTION SHARING & ADAIN

In our framework:

- **Shared Attention** enables **Consistent Style Alignment** across **Multiple Images** during the diffusion process. Each image in the set attends not only to itself but also to a Reference Image (in this case the first one in the batch.
- **Adaptive Instance Normalization (AdaIN)** has the aim to **adjust** the **target image's queries** and **keys** based on the reference image's statistics, by:
  - **Reducing Content Leakage**
  - **Maintaining diversity** across generated images

$$Sh\_Attention(Q_t, K_{ref}, V_{ref})$$

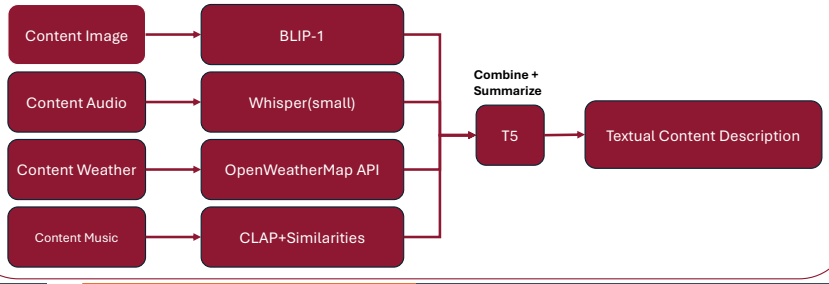
$$+$$

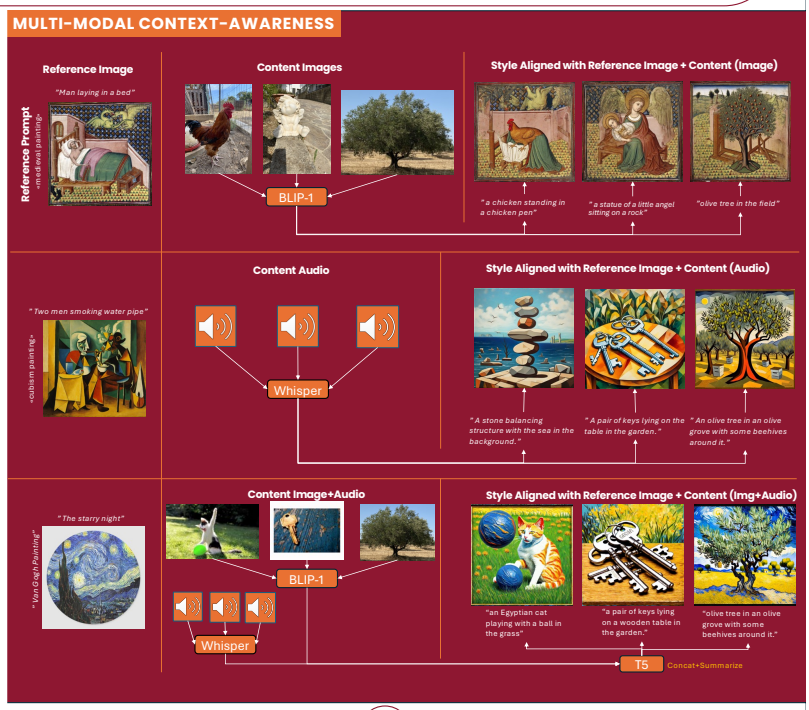
$$Q'_t = AdaIN(Q_t, Q_{ref})$$
$$K'_t = AdaIN(K_t, K_{ref})$$

Where:

$$AdaIN(x, y) = \sigma(x) * \left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

MULTI-MODAL CONTEXT-AWARENESS





MULTI-REFERENCE STYLE

- **Style-Aligning** the generated images from **Multiple Reference Images**.
- Ability to also **weight** the Style Contribution.
- Image **style combination** L_i occurring in the **Image Latent Space**.
- **Blending** Options: **Linear Weighted Image** and **SLERP.**

$$L_i = VAE_{Encoder}(T_i) \cdot Scaling_{factor}$$

$$L_{LinearBlending} = \sum_{i}^{n} w_i \times L_i$$

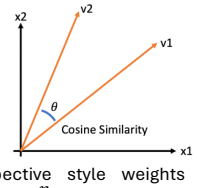
$$Slerp(t, v_0, v_1) = \frac{\sin((1-t)\cdot\omega)}{\sin(\omega)} \cdot v_0 + \frac{\sin(t\cdot\omega)}{\sin(\omega)} \cdot v_1$$

$$\omega = \arccos\left(\frac{v_0 \cdot v_1}{\|v_0\| \cdot \|v_1\|}\right)$$

$$L_{SLERPBlending} = Slerp(w_i, L_{LinearBlending}, L_i)$$

WEIGHTED MULTI-STYLE DINO VIT-B/8

- Need for a robust evaluation method in **multi-style image generation**.
- Traditional metrics fall short when blending multiple styles, since don't capture the combined influence of different references.
- **Introducing** a **New Metric** called **Weighted Multi-Style DINO VIT-B/8.**
- Cosine similarity S_{ri} of a generated image embedding I_i to each reference style image embedding $I_{r1}, I_{r2}, \ldots, I_{rn}$, applies the respective style weights $w_1, w_2, \ldots, w_n$, and provides a final composite score:



$$WMS_{DINO-ViT-B/8} = \sum_{i=1}^{n} S_{ri} \cdot w_i$$

RESULTS & FUTURE WORKS

Linear Weighted Blending

Medieval/Picasso Style	$WMS_{DINO-ViT-B/8}$
1 / 0	0.32538
0.25 / 0.75	0.39226
0.5 / 0.5	0.35965
0.75 / 0.25	0.38044
0 / 1	0.44461

Spherical Weighted Blending

Medieval/Picasso Style	$WMS_{DINO-ViT-B/8}$
1 / 0	0.38842
0.25 / 0.75	0.44753
0.5 / 0.5	0.31265
0.75 / 0.25	0.39800
0 / 1	0.48312

- Linear Attention
- Audio-Tone Voice Recognition