Z-SASLM: Zero-Shot Style-Aligned SLI Blending Latent Manipulation

Alessio Borgi, Luca Maiano, Irene Amerini
Sapienza University of Rome, Italy

Project Page

Connect with Me

CVEU - *AI for Creative Visual Content Generation Editing and Understanding*

Introduction

Style-Alignment across a set of generated images

Challenges

- Requires fine-tuning.
- Limited to single style-reference in Image conditioning.
- Content and style disentanglement with multiple styles.

Proposed Solution

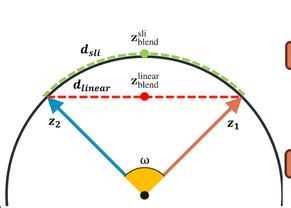
- **Zero-Shot Multi-Reference-Weighted Style-Alignment** in **Image generation** through **Spherical Linear Interpolation Blending** via **Latent Manipulation**.

Multi-Reference-Weighted SLI Style Blending

- **Latent-Space manipulation** technique.
- Linear Interpolation applied to latent vectors leads to artifacts and inconsistencies.
- z_{blend} used as conditioning latent within diffusion process.
- **Spherical Linear Interpolation** follows the Geodesic in the hypersphere.

$$SLI(t, z_1, z_2) = \frac{\sin((1-t)\cdot\omega)}{\sin(\omega)} \cdot z_1 + \frac{\sin(t\cdot\omega)}{\sin(\omega)} \cdot z_2$$

$$\omega = \arccos\left(\frac{v_0 \cdot v_1}{\|v_0\| \cdot \|v_1\|}\right), \qquad t = \frac{w_2}{w_1 + w_2}$$



Style-Alignment: Shared Attention Layer

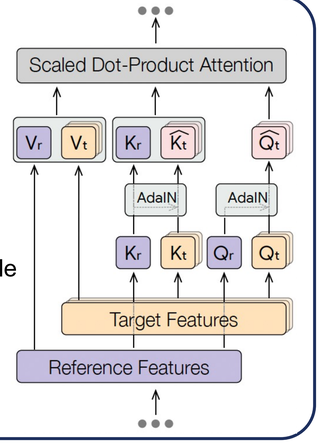
- **AdaIN** – aligns generated image feature stats with reference style image.
- Apply AdaIN over target queries Q_t and keys K_t using reference queries Q_r and keys Q_k.

$$\hat{Q}_t = AdaIN(Q_t, Q_r) \ , \quad \hat{K}_t = AdaIN(K_t, K_r)$$

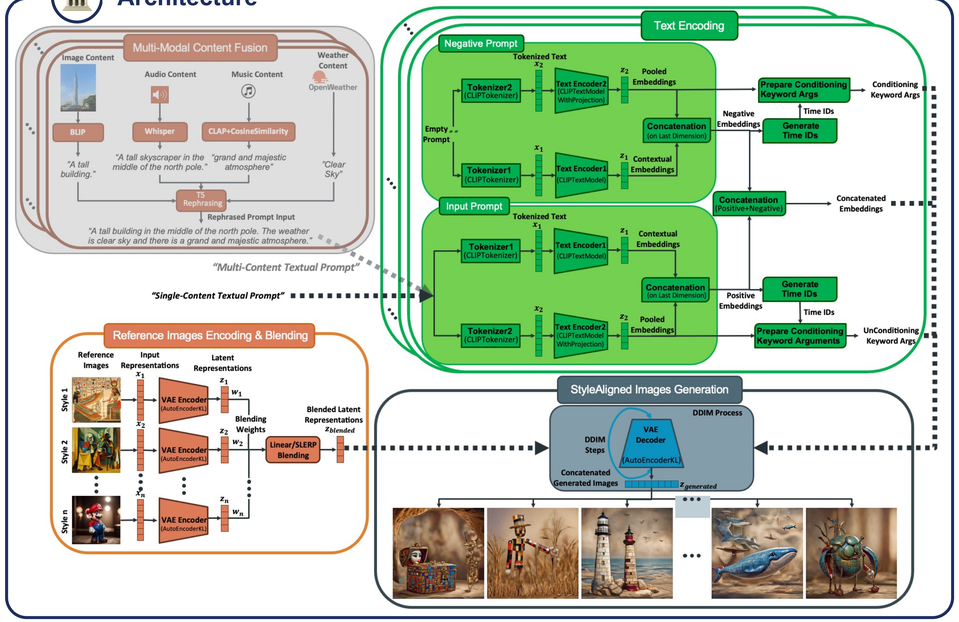
$$AdaIN(x, y) = \sigma(x) * \left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

- **Shared Attention** – propagates style info across multiple
- generated images.
- Target features updated by both target values V_t and reference values V_r .

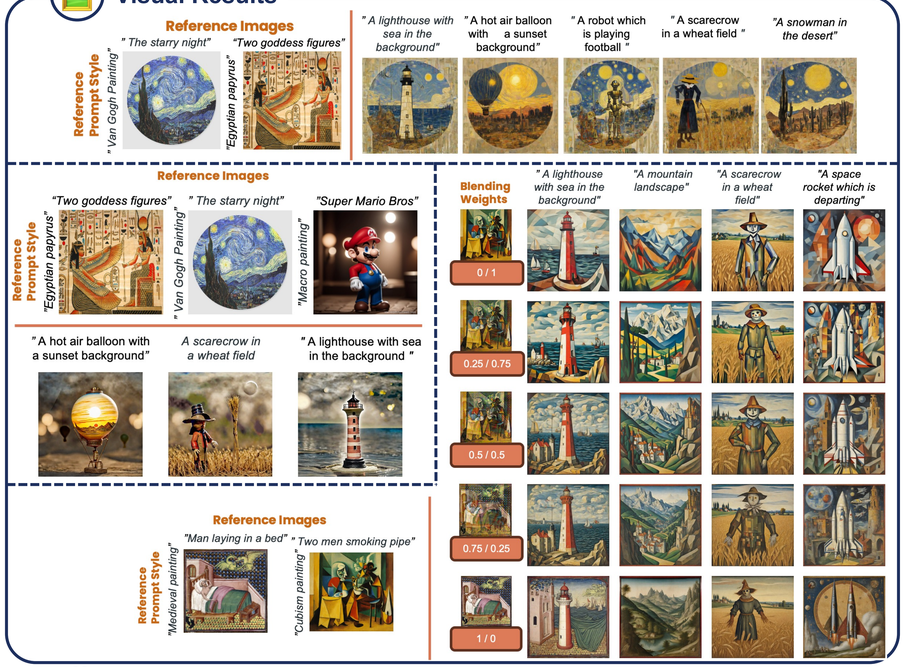
$$ShAttention(\hat{Q}_t, K^T_{rt}, V_{rt}) \ , \quad K_{rt} = \begin{bmatrix} K_r \\ \hat{K}_t \end{bmatrix} \quad V_{rt} = \begin{bmatrix} V_r \\ V_t \end{bmatrix}$$



Architecture



Visual Results



Results & Weighted Multi-Style DINO VIT-B/8

- Cosine-Similarity over $WMS_{DINO\ VIT-B/8}$ feature embeddings to assess generated image vs. reference image style-alignment.
- $CLIP_{score}$ to evaluate image-text alignment.

Style Weights	Linear (StyleGAN2-ADA[21]-adapted)				Z-SASLM (Ours)			
$\{w_{med}, w_{cub}\}$	MS_{med}	MS_{cub}	$WMS_{DINO\text{-}VIT\text{-}B/8}$	$CLIP_{score}$	MS_{med}	MS_{cub}	$WMS_{DINO\text{-}VIT\text{-}B/8}$	$CLIP_{score}$
$\{0, 1\}$*	-	0.47552	0.47552	0.30280	-	0.47552	0.47552	0.30280
$\{0.15, 0.85\}$	0.32466	0.42683	0.41151	**0.31534**	0.32595	0.47072	**0.44900**	0.31049
$\{0.25, 0.75\}$	0.35550	0.42250	0.40575	0.31420	0.33046	0.45447	**0.42347**	**0.31657**
$\{0.5, 0.5\}$	0.34905	0.37881	0.36393	0.29232	0.36150	0.42156	**0.39153**	**0.31434**
$\{0.75, 0.25\}$	0.35798	0.38327	**0.36430**	0.31752	0.34648	0.35099	0.34760	**0.31911**
$\{0.85, 0.15\}$	0.35513	0.40860	0.36315	**0.32381**	0.36513	0.38286	**0.36779**	0.31499
$\{1, 0\}$*	0.29891	-	0.29891	0.30570	0.29891	-	0.29891	0.30570