AI for Visual Art Workshop and Challenges

Alfie: Democratising RGBA Image Genearion With No $$$

Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, Rita Cucchiara

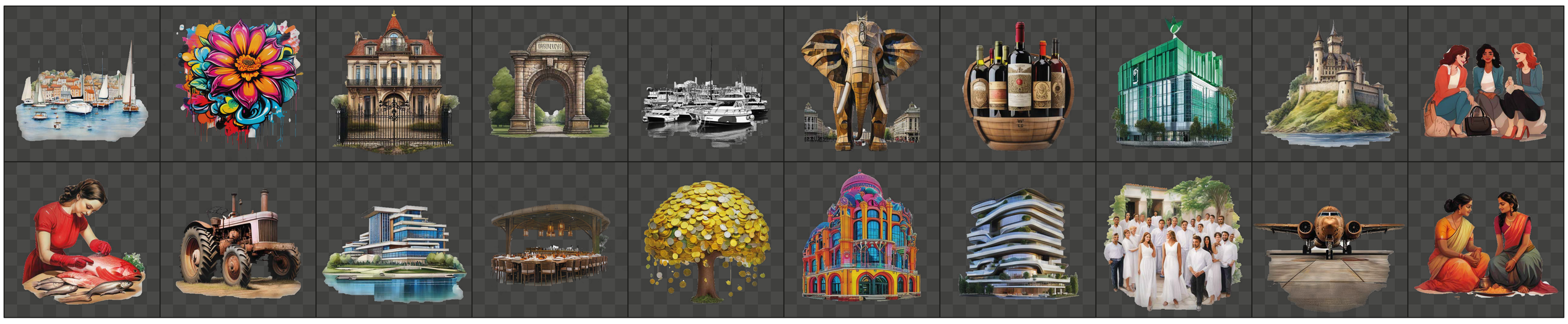
name.surname@unimore.it

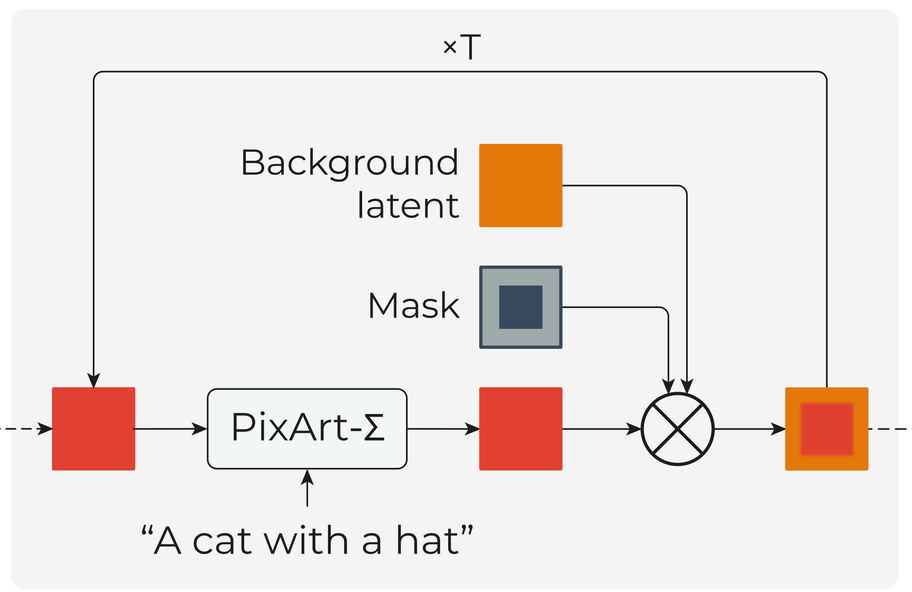
AImage Lab

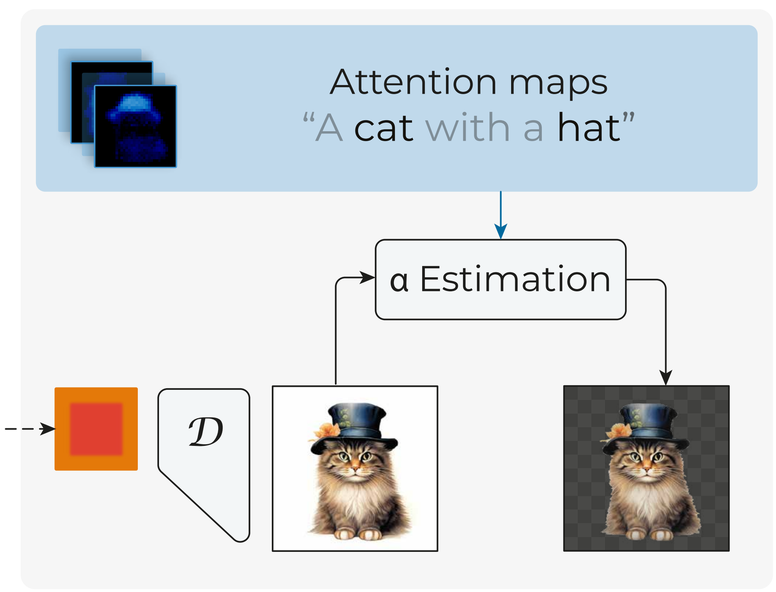
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

An inference-time pipeline to obtain prompt-guided RGBA illustrations from a Diffusion Model trained for RGB images







Prompt-guided RGBA Illustrations should:

- Respect the prompt
- Be fully contained in the canvas
- Have a precise **α** channel

We obtain them by generating a suitable RGB image with PixArt-**Σ** [1] and then estimating its **α** channel

We define a squared mask **m**, covering the inner area of the canvas

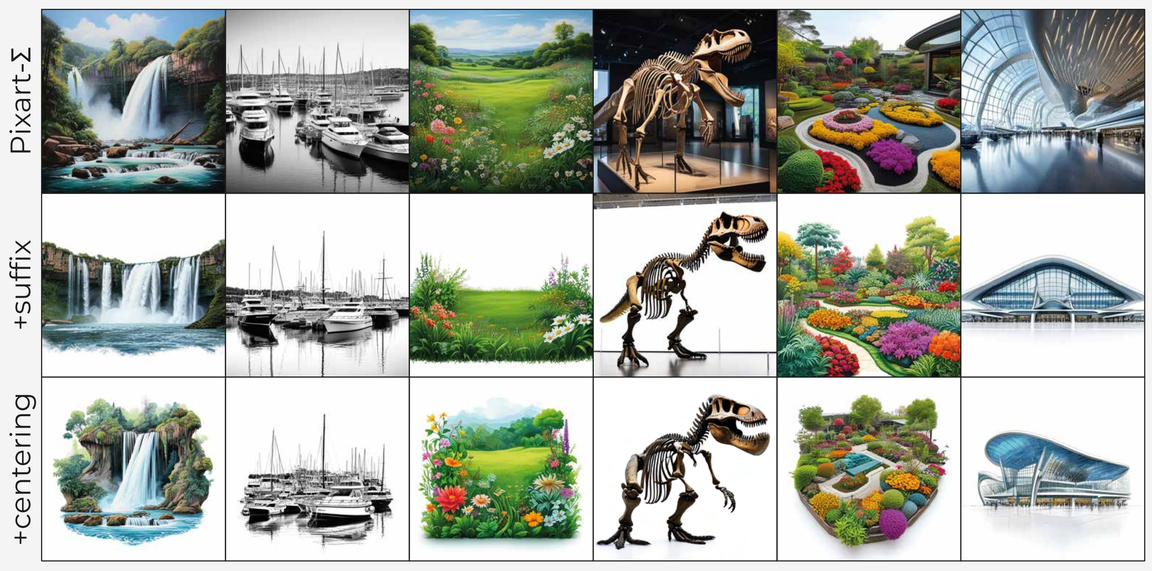
We jointly generate

- The illustration (described by the *input* prompt) inside the mask
- The background (described by the prompt "*A white background*") outside the mask

We blend them together at each denoising step:

$$\mathbf{x}_{t-1} = \mathbf{x}_{t-1,fg} \cdot \mathbf{m} + \mathbf{x}_{t-1,bg} \cdot (1 - \mathbf{m})$$

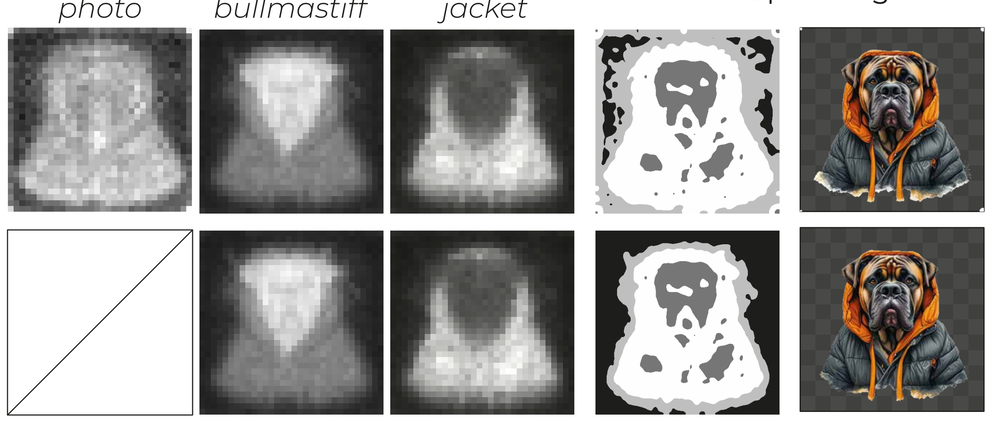
	empty-a	CLIP-S
Pixart-Σ	3.33	31.29
+suffix	53.10	30.79
+centering	96.50	30.08



Starting from the centered RGB illustration on uniform white background...

We combine the cross-attention maps relative to the (interesting) nouns in the prompt

NB: Not all prompt nouns are relevant



Prompt: *A photo of a bullmastiff with a jacket*

We extract the subject information from the self-attention maps [2]

$$\hat{\alpha} = \overline{CA}_{fg} + \overline{\mathcal{FF}}_{(\overline{CA}_{fg})}$$

We apply GrabCut by exploiting the 4-value quantized candidate foreground maps

We adjust the opacity through an hyperparameter: $\hat{\alpha}' = \min(1, (1 + k)\hat{\alpha})$

Alfie gives better-matted illustrations compared to a pipeline entailing RGB image generation + matting (with quantized attention maps as trimaps) [3]

Alfie's illustrations can be simply superimposed to or blended with a desired background [4]

REFERENCES

[1] Chen, J. et al.: PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4KText-to-Image Generation. ECCV (2024)

[2] Chowdhury, R.D.: DAAMImage2Image: Extension of DAAM for Image Self-Attetion in Diffusion Models. GitHub

[3] Yao, J. et al.: ViTMatte: Boosting image matting with pre-trained plain vision transformers. Inform. Fusion (2024)

[4] Sarukkai, V. et al.: Collage Diffusion. WACV (2024)