CosmicMan

- □□□□□CosmicMan: A Text-to-Image Foundation Model for Humans
- □□□□□https://arxiv.org/abs/2404.01294
- CVPR 2024

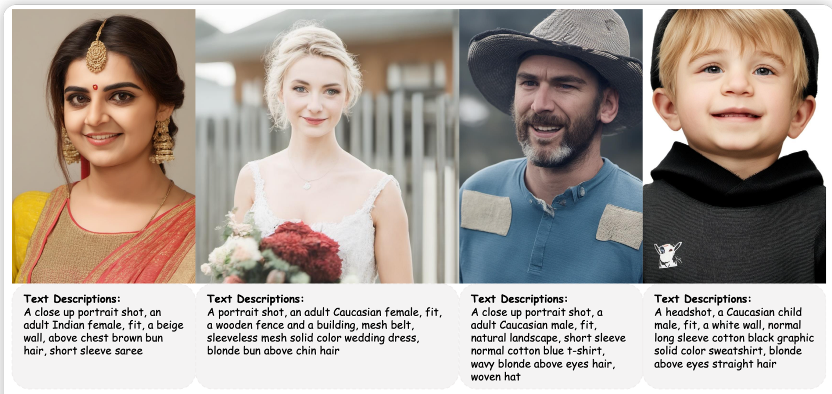

Figure 1. **CosmicMan.** High-fidelity images generated by our proposed human-specialized text-to-image foundation model CosmicMan. The results are with meticulous appearance, reasonable structure, and precise text-image alignment with detailed dense descriptions. More results are presented in the project page: https://cosmicman-cvpr2024.github.io/.

□□□□□CosmicMan□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□Annotate Anyone□□CosmicMan-HQ 1.0□□□□□6M□□□□□□□□□□□□□□□□□□□□1488*1255□□□□□115M□□□□□□□□□□□□□□

□□□□□□□□□□□□SD□□□□□Daring□□□

□□□□□□□□□□□

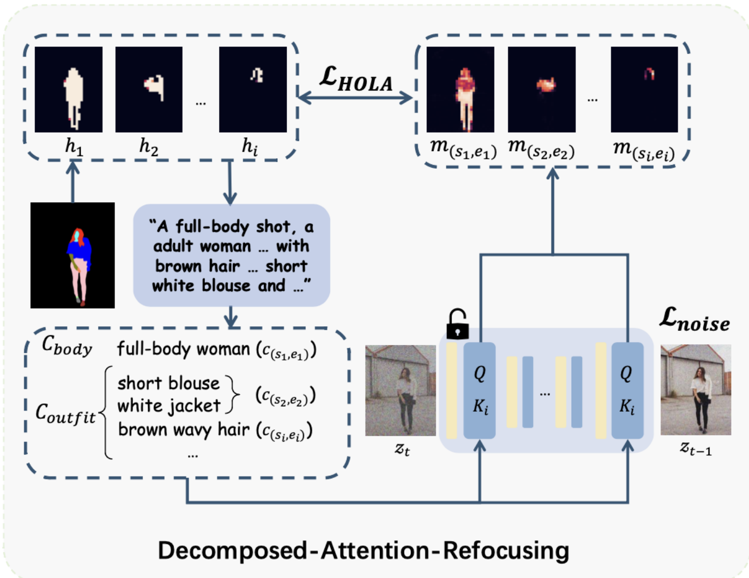

Figure 4. **Daring Training Framework.** It includes two parts: (1) data discretion for decomposing text-human data into fixed groups that obey human structure; (2) a new loss – HOLA, to enforce the cross-attention features actively response in proper spatial region with respect to the scale of body structure and outfit arrangement.

□□□□□□□□□□□□□□□□caption□□□□□□□□□□□□□□□□□□□□□□□□□□□mask□□□□caption□□□□□□□□□□□□□□□□□□□□□caption□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□HOLA□□□

$$\mathcal{L}_{\text{HOLA}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=s_i}^{e_i}\|m_j - h_i\|_2^2 + \left\|\frac{1}{e_i - s_i}\sum_{j=s_i}^{e_i}(m_j) - h_i\right\|_2^2\right)$$

□□

- □□□CosmicMan-HQ 1.0
- □□□FID□□□□□□□□□□CLIP□□□□□□□□□□□□□User Study
- □□□32 A100/bs64
- □□□https://cosmicman-cvpr2024.github.io
