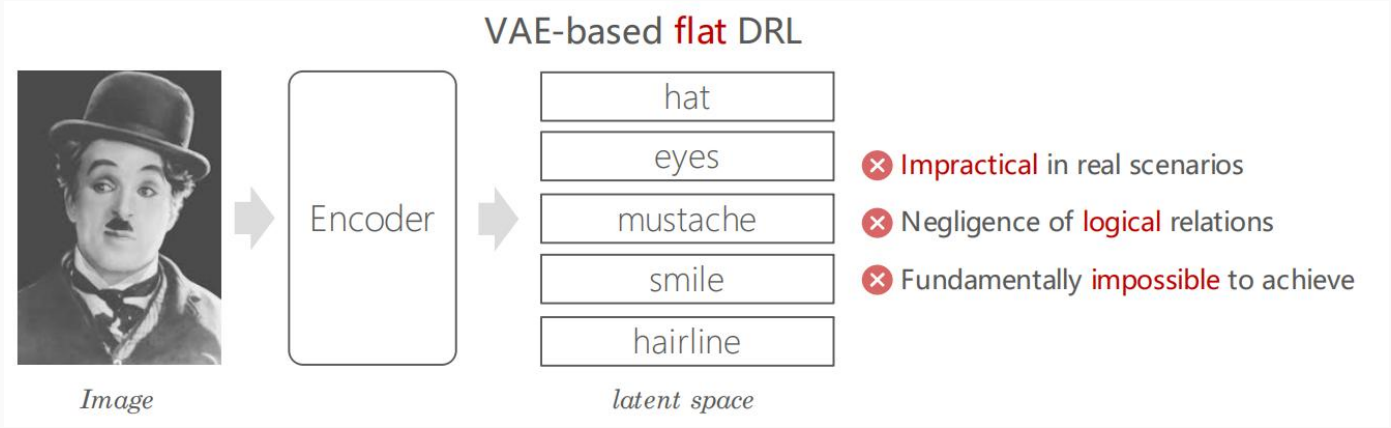
Graph-based Unsupervised Disentangled Representation Learning via Multimodal Large Language Models

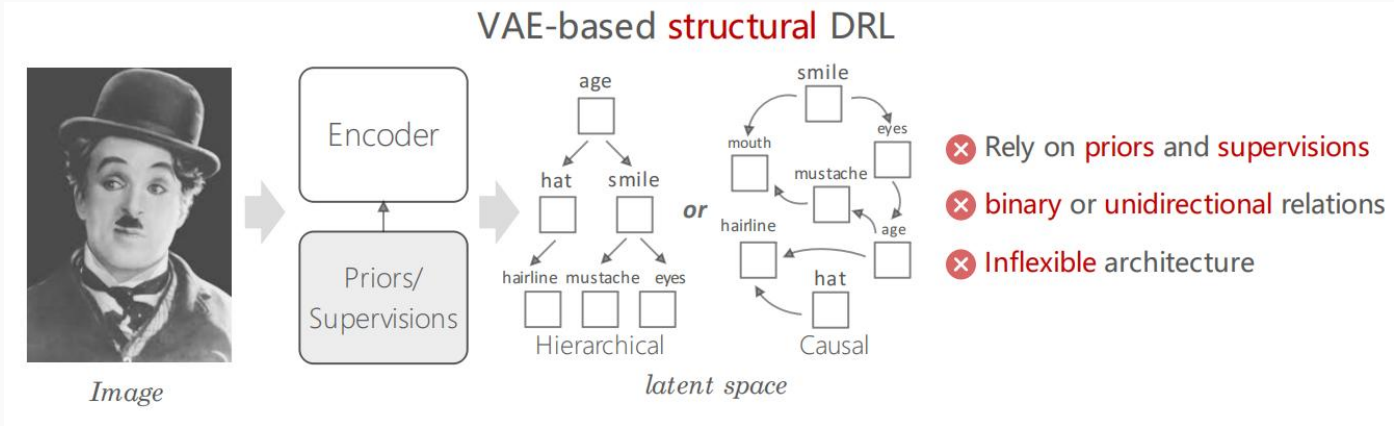
Baao Xie, Qiuyu Chen, Yunnan Wang, Zequn Zhang, Xin Jin*, Wenjun Zeng

by Qiuyu Chen

2024.11.05

Motivation:





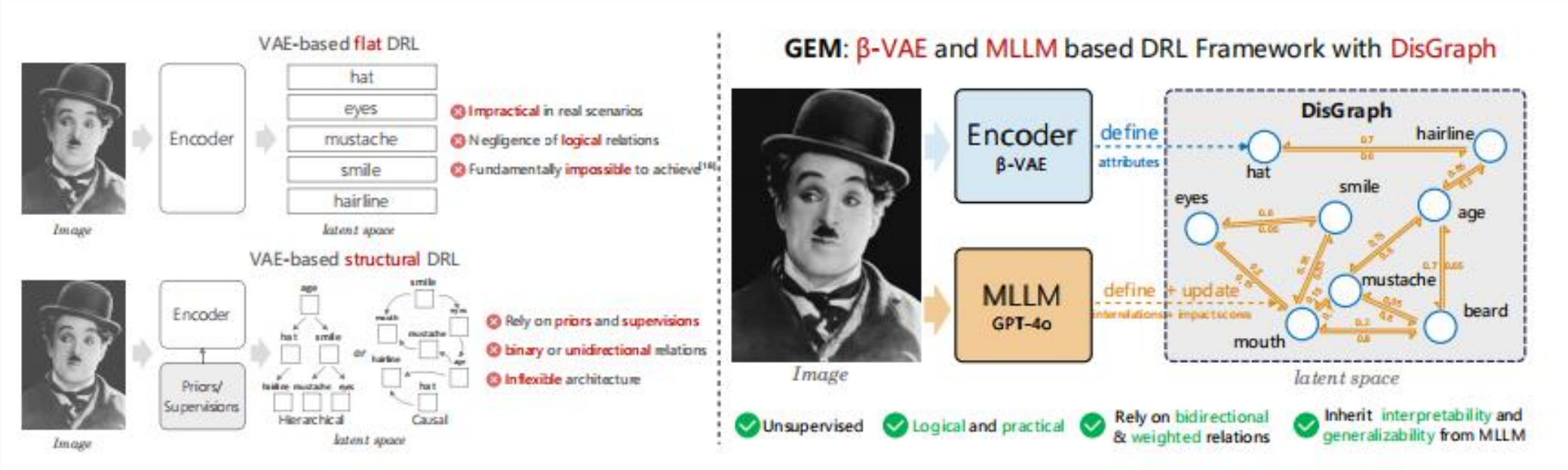
VAE-based flat DRL

- Assumptions about the real world are too idealistic
- Ignoring the connection between attributes
- Poor quality of generated and reconstructed images

VAE-based structural DRL

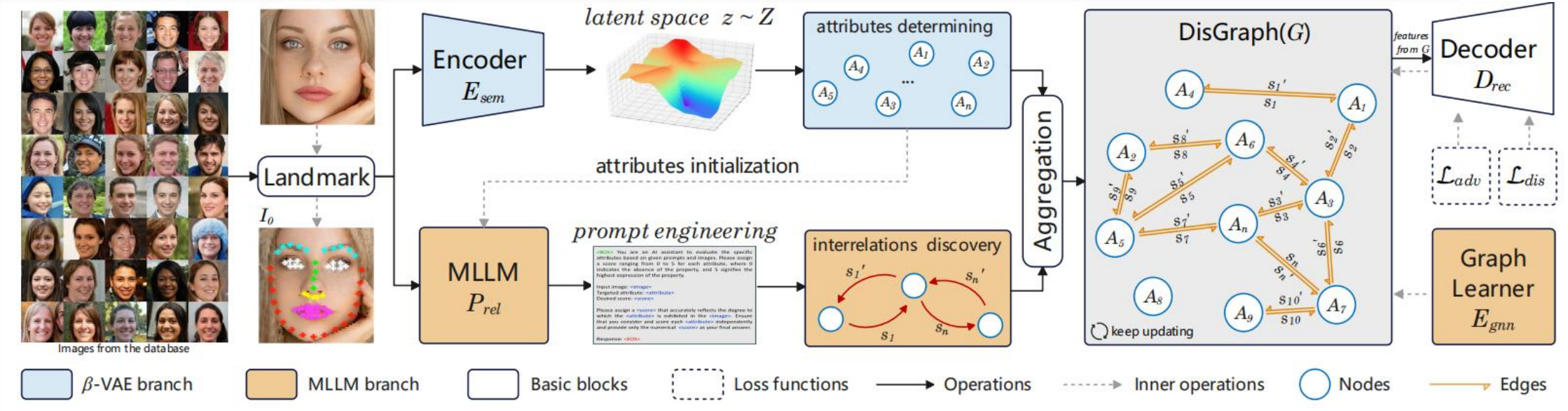
- Heavily dependent on supervision and prior
- Modeling relationships between attributes is too simplistic
- Poor generalization

Contribution：



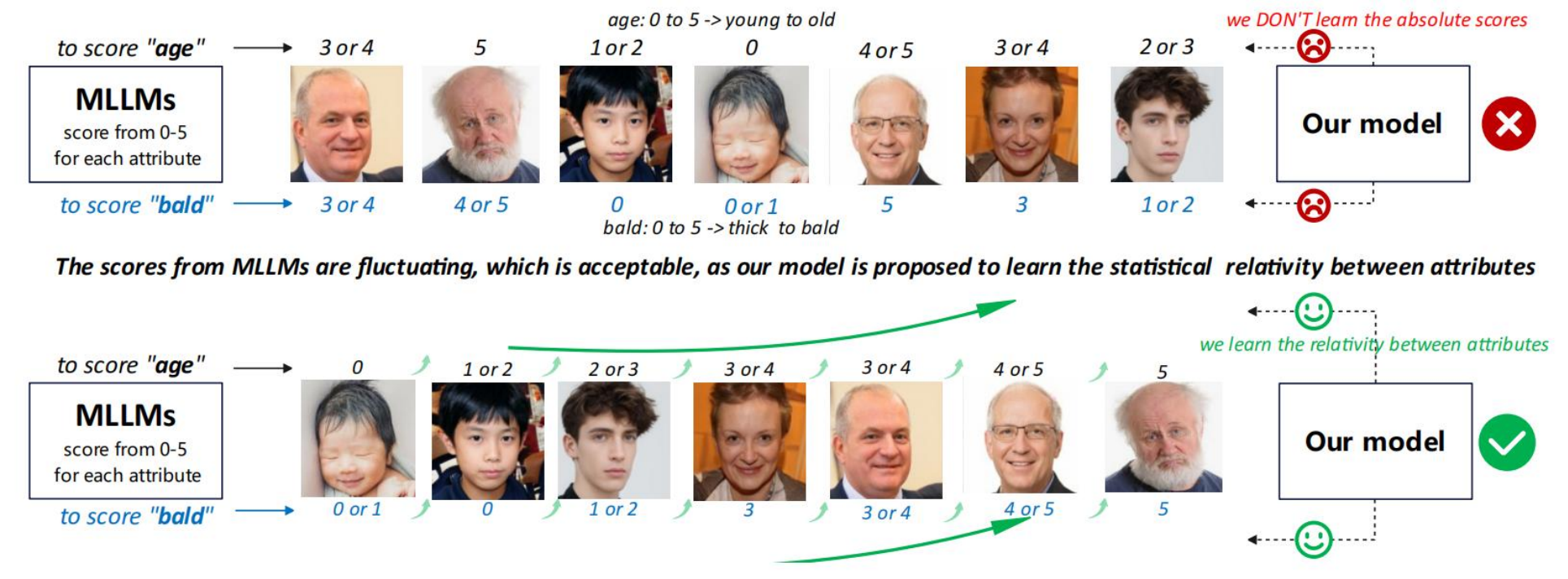
- the framework is fully unsupervised
- superior performance on fine-grained and relation-aware disentanglement
- Ability to work with practical scenarios
- superior interpretability and generalizability

Pipeline of GEM



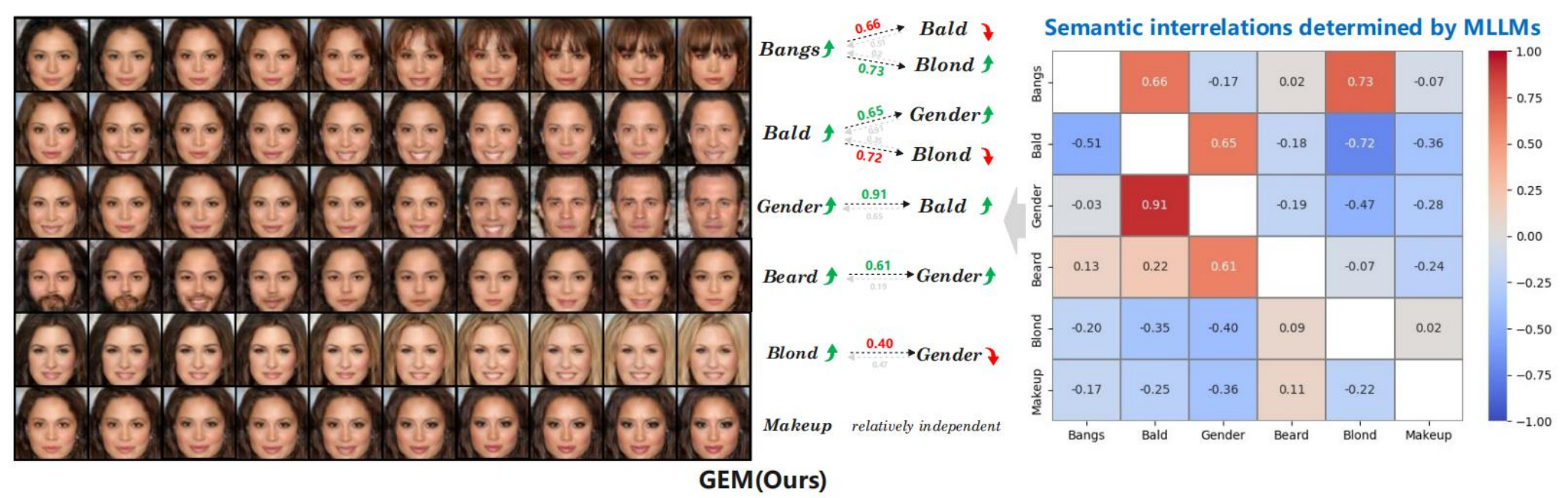
- Attribute extraction module: It is used to disentanglethe target attributes and provide the initialization of the attributes for the associated prediction module.
- Association prediction module: mining, sorting and weighting the association between attributes.
- Bidirectional weighted explicit graph: Encodes the above semantic information, the representation attributes are nodes, the correlation is represented as edges, and the similarity coefficient is represented as weights.

MLLM-based Interrelation Discovery Branch



Our aim is using the commonsense knowledge behind MLLMs to equip GEM with ability of interrelations discovery, where a certain degree of fluctuations on absolute scores are acceptable.

Experimental results



GEM effectively realizes fine-grained and relation-aware representation disentanglement through integrated disentangled representation learning and MLLMs. Each row in facial images corresponds to the traversal results on a specific attribute, as indicated adjacent to the images.

Experimental results

GEM surpasses baseline models in reconstruction quality on the datasets of CelebA, LSUN-horse, and LSUN-bedroom. The results indicate that GEM outperforms both typical unsupervised and supervised approaches in terms of reconstruction quality.

Method	CelebA		LSUN-horse		LSUN-bedroom	
	FID ↓	KID $\times 10^3$ ↓	FID ↓	KID $\times 10^3$ ↓	FID ↓	KID $\times 10^3$ ↓
VAE [13]	53.3 ± 0.6	51.4 ± 0.4	172.8 ± 1.7	181.7 ± 2.1	195.8 ± 4.1	226.4 ± 5.4
β-VAE [13]	136.2 ± 1.6	107.0 ± 2.7	272.4 ± 3.2	294.2 ± 5.3	288.1 ± 5.7	225.7 ± 6.0
β-TCVAE [14]	139.1 ± 0.8	113.2 ± 4.1	173.0 ± 4.8	217.35 ± 9.2	191.0 ± 5.0	179.2 ± 7.4
FactorVAE [41]	134.5 ± 0.3	92.0 ± 0.5	248.5 ± 5.5	155.3 ± 3.7	235.7 ± 3.2	172.8 ± 3.9
DEAR [27]	70.7 ± 0.3	52.6 ± 0.1	136.4 ± 1.6	113.7 ± 0.9	177.6 ± 3.5	157.8 ± 2.3
GEM (Ours)	**46.0 ± 0.1**	**48.3 ± 0.2**	**101.0 ± 1.1**	**65.5 ± 1.7**	**125.4 ± 1.2**	**76.1 ± 1.1**

Thank you for your watching!