Image-guided topic modeling for interpretable privacy classification

Alina Elena Baia[1] and Andrea Cavallaro[1,2]

[1]Idiap Research Institute, [2]École Polytechnique Fédérale de Lausanne

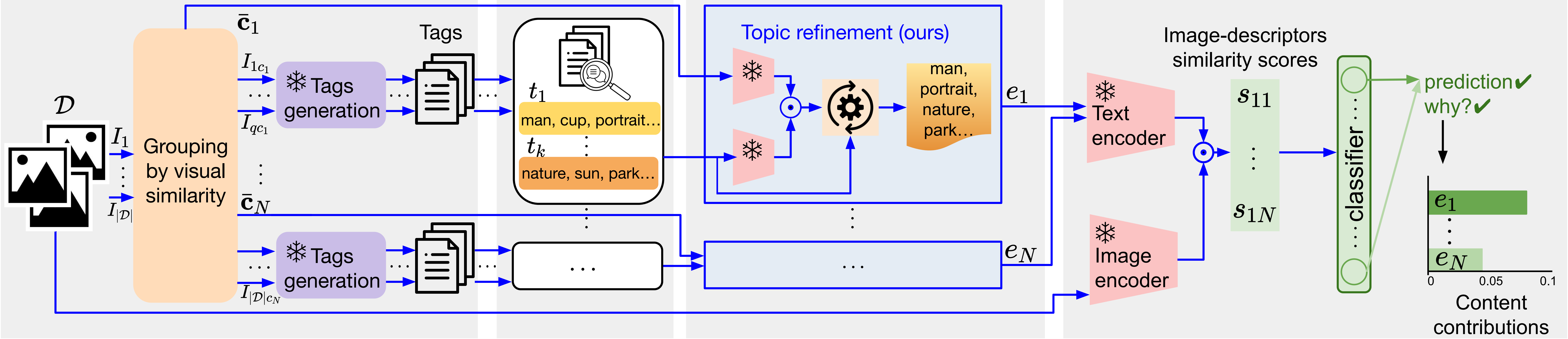
Introduction

- Image privacy: a complex and contextual task, challenging even for LLMs/VLMs
- Human-centric approach to help users understand privacy risks
- Novel multimodal framework that enables the learning of a classifier whose decisions can be interpreted using natural language

VLMs limitations in privacy

VLM prompt: *Does this image contain privacy sensitive content? If yes, which content is private?*

CogVLM-17B: *Yes, this image contains privacy sensitive content. The picture shows a bee in a black background with its eyes and mouth covered… This information might violate the privacy of the bee or any other living being…*



$\mathcal{D}$

I_1

$I_{|\mathcal{D}|}$

Grouping by visual similarity

$\bar{\mathbf{c}}_1$

I_{1c_1}

I_{qc_1}

$\bar{\mathbf{c}}_N$

$I_{|\mathcal{D}|c_N}$

❄ Tags generation

❄ Tags generation

Tags

t_1

man, cup, portrait…

t_k

nature, sun, park…

Topic refinement (ours)

man, portrait, nature, park…

e_1

e_N

❄ Text encoder

❄ Image encoder

Image-descriptors similarity scores

s_{11}

s_{1N}

… classifier …

prediction✔

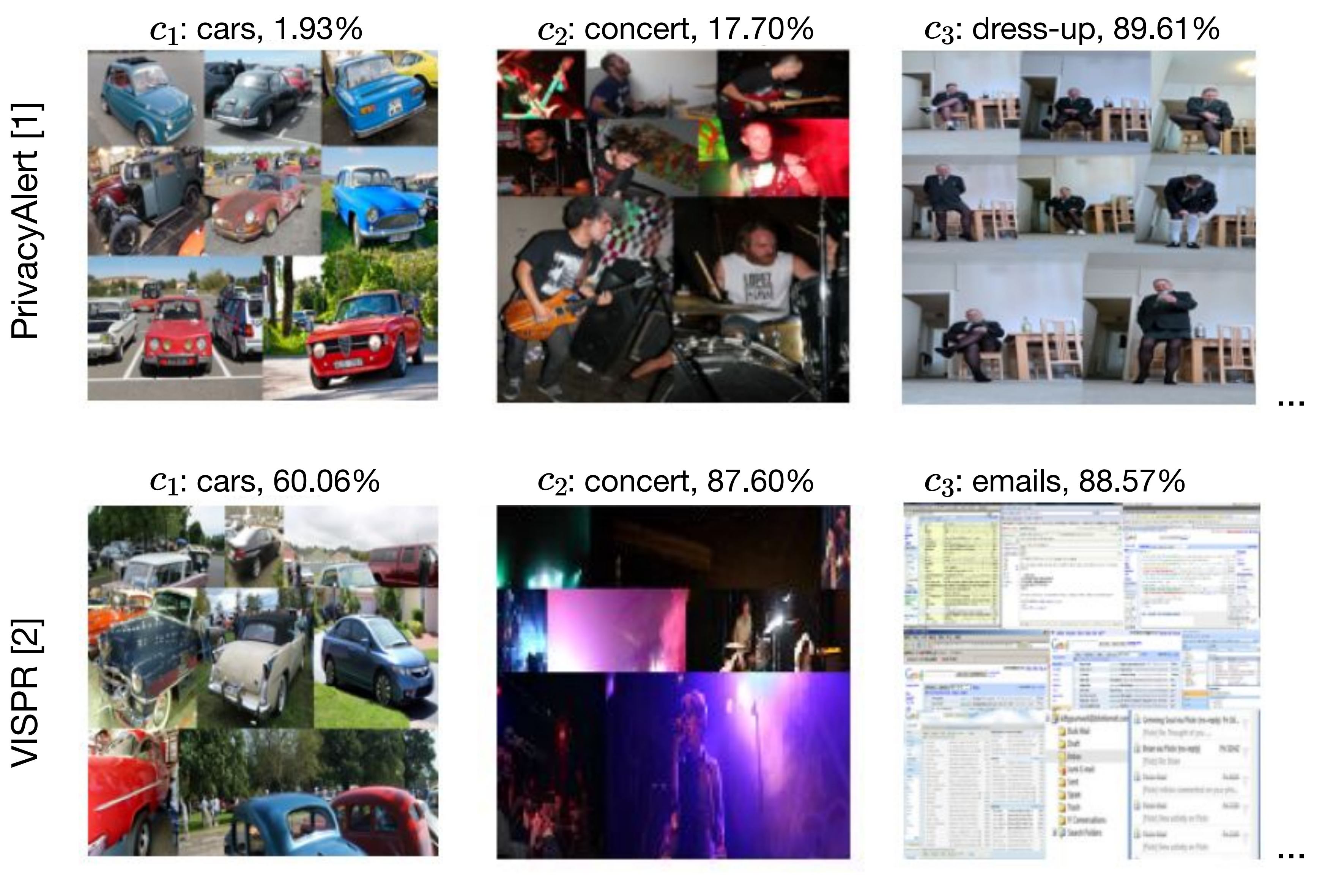
why?✔

e_1

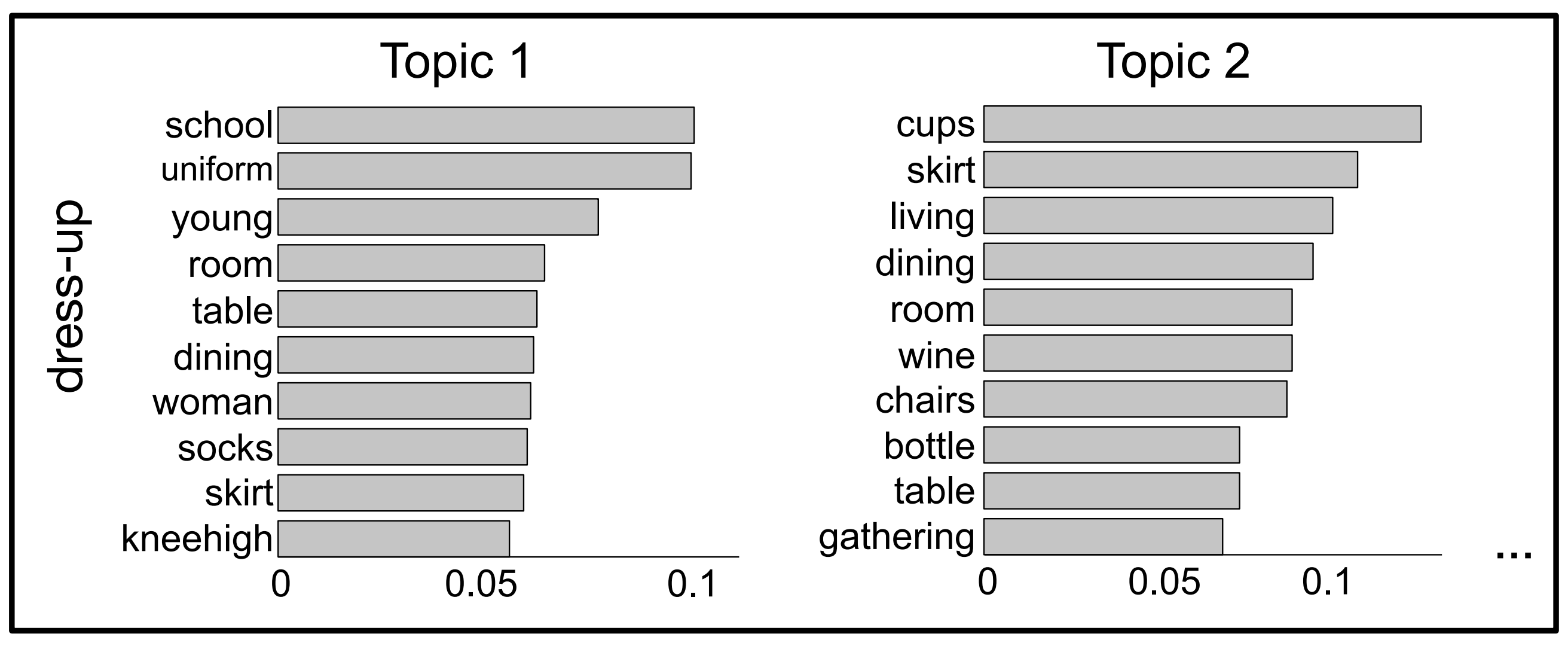
e_N

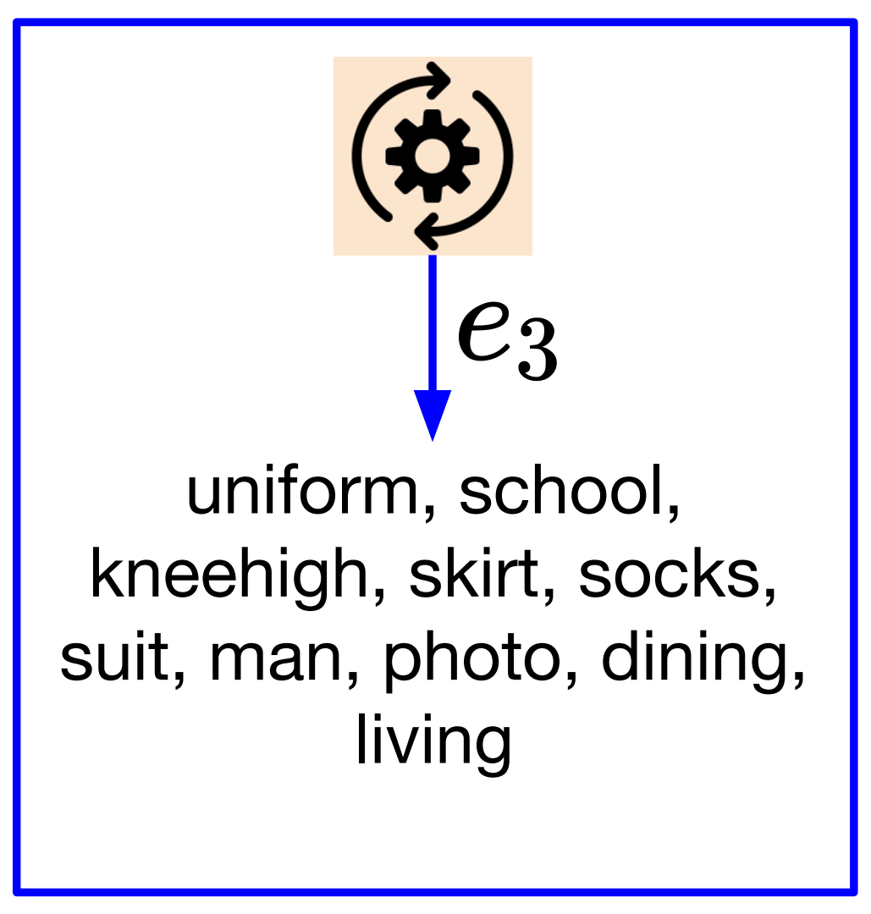
0 0.05 0.1

Content contributions



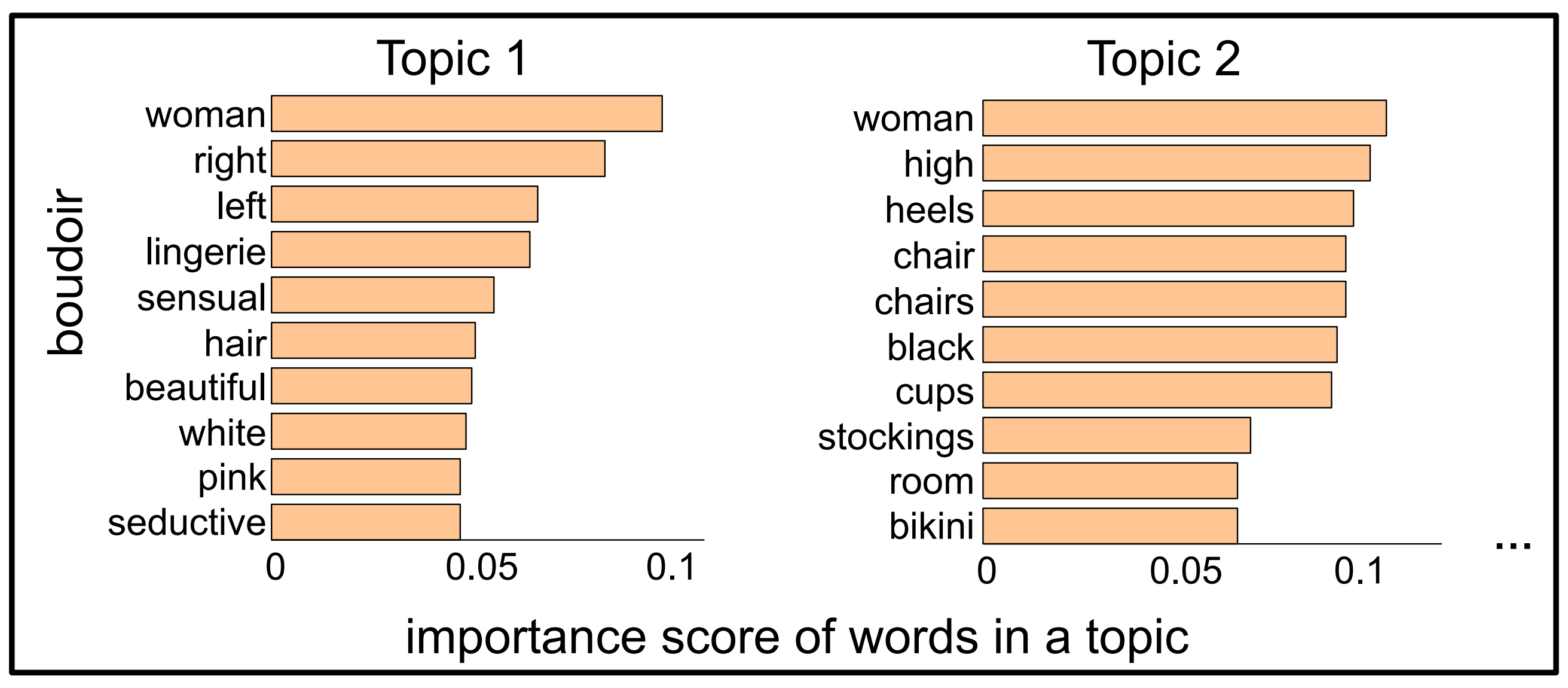
Clusters (privacy levels %) obtained for image privacy datasets. Note that the same content (i.e. cars, concert) was annotated differently in the two datasets.

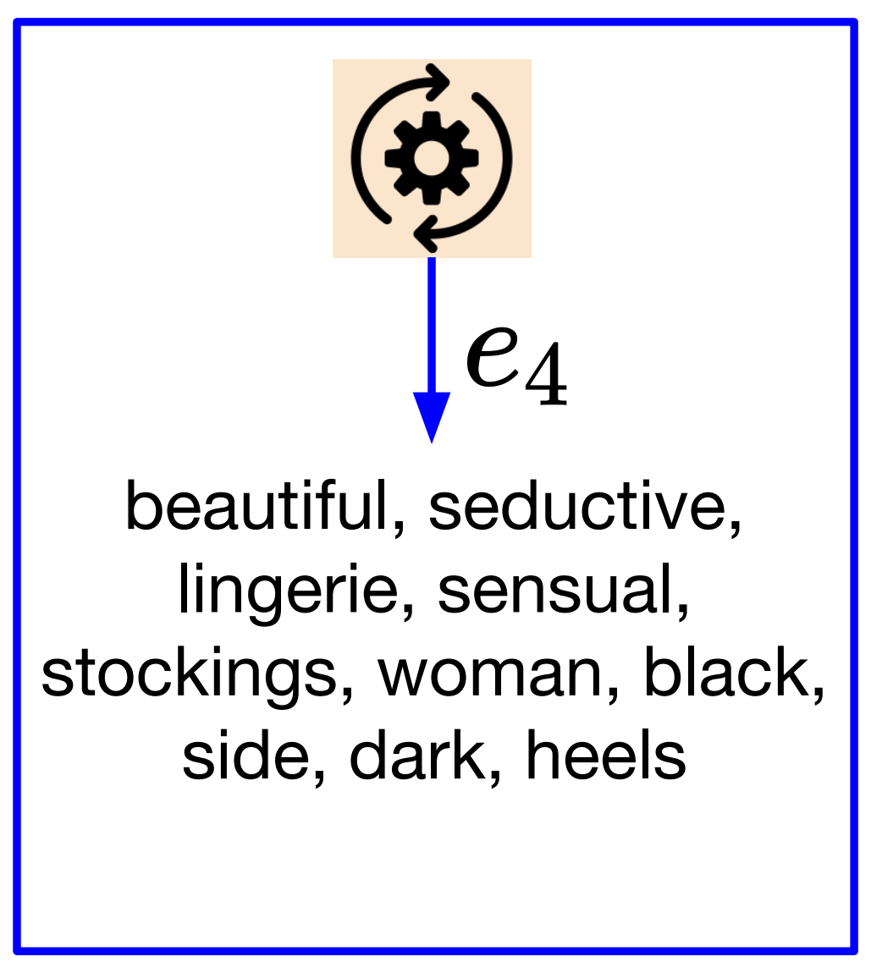




e_3

uniform, school, kneehigh, skirt, socks, suit, man, photo, dining, living

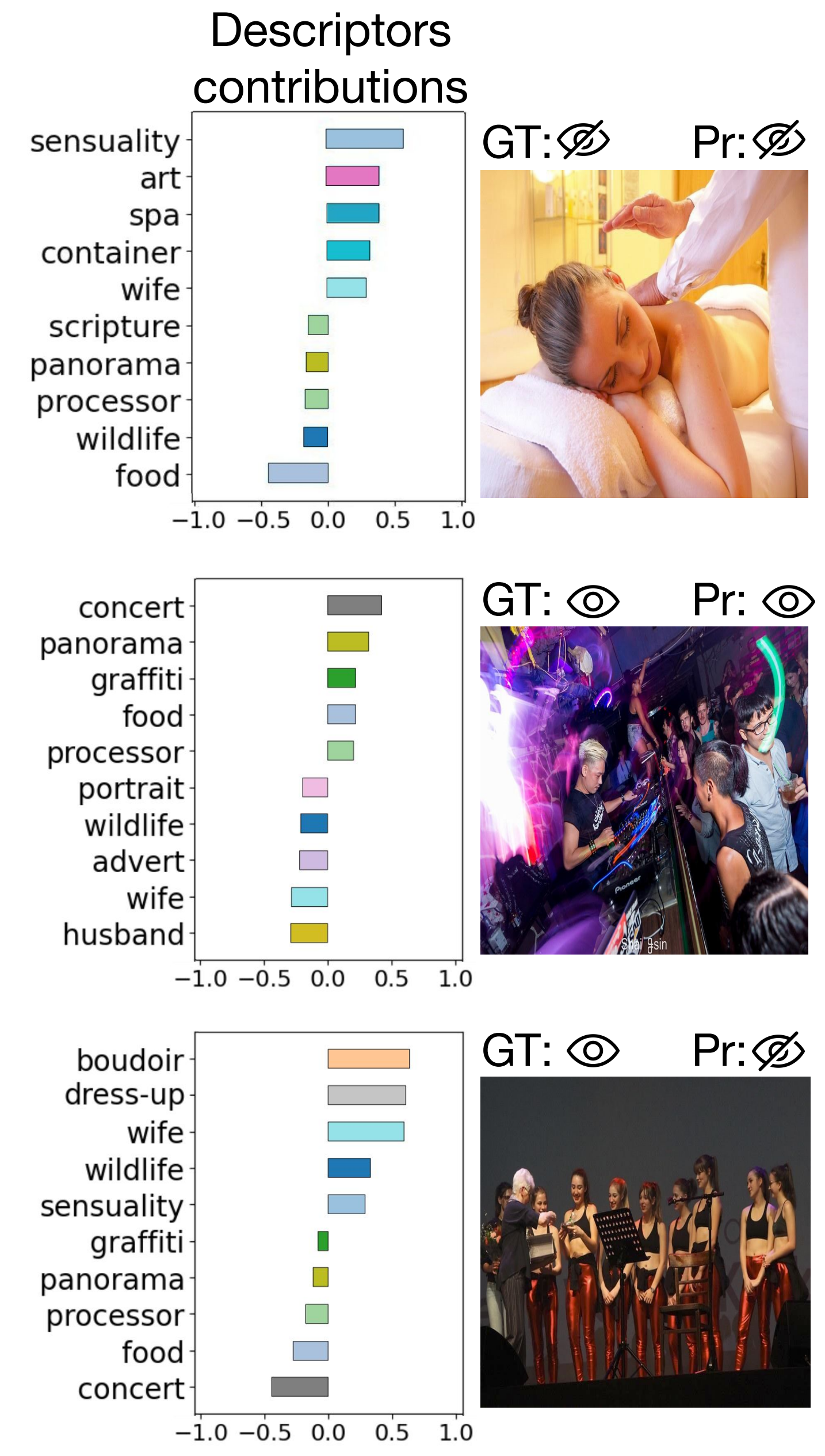




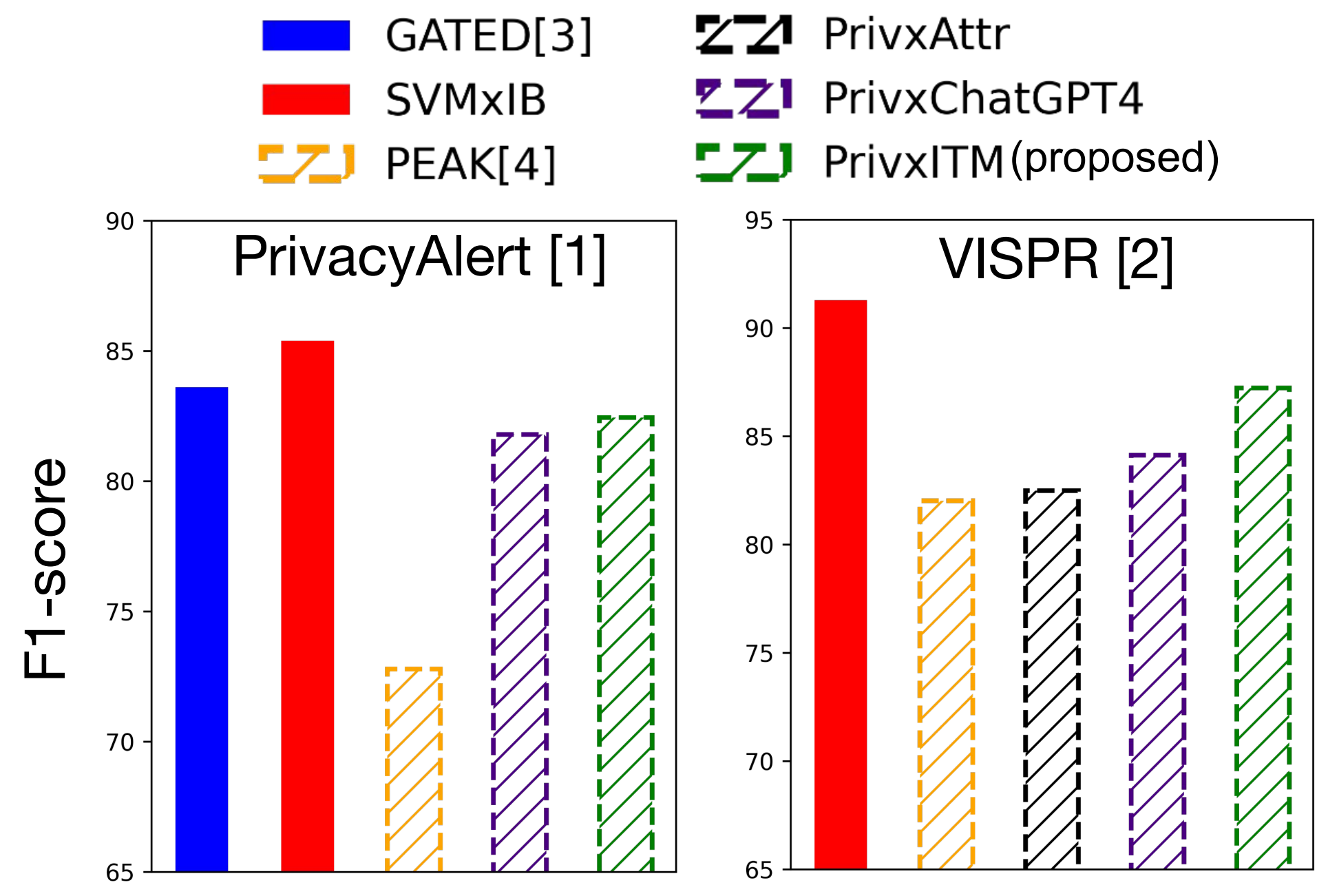
e_4

beautiful, seductive, lingerie, sensual, stockings, woman, black, side, dark, heels

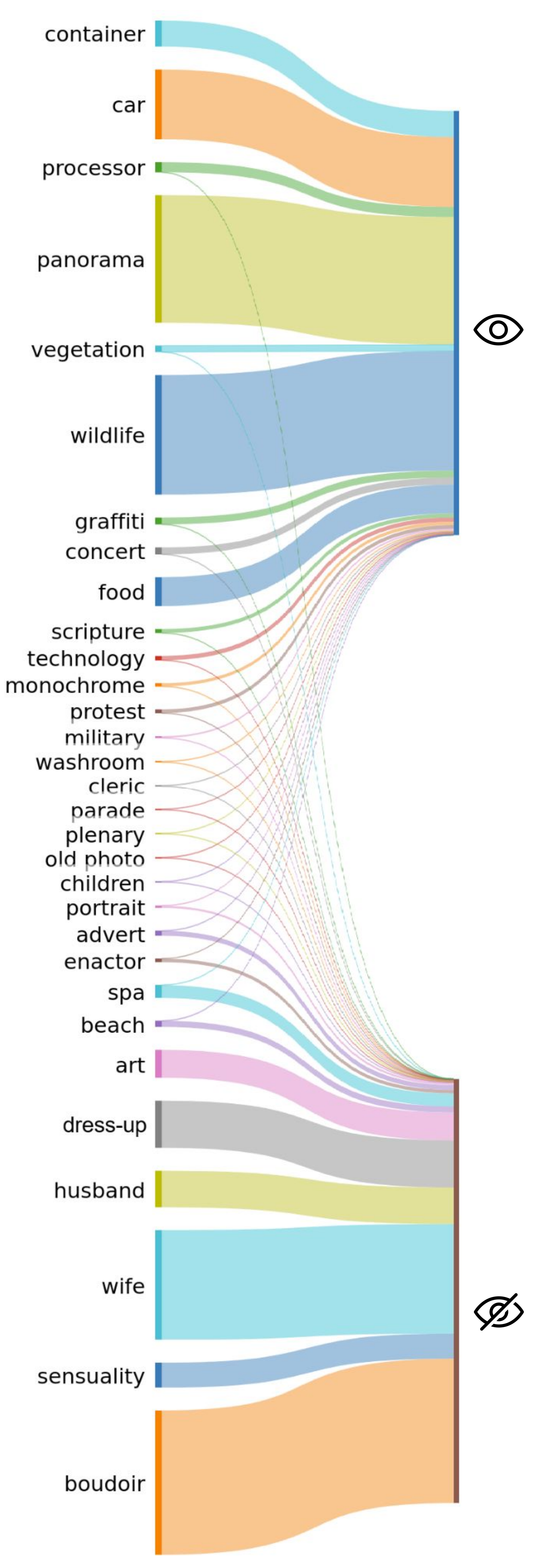
- Topics representation discovered within the image clusters *dress-up* and *boudoir* of PrivacyAlert.
- Topic refinement removes objects that are often hallucinated such as cups and chairs, and generates the cluster descriptors e_i.



Interpretations of Priv×ITM predictions for PrivacyAlert using the top-5 positive and negative descriptor contributions



Classification results (F1-scores). PrivxAttr: human annotated attributes PrivxChatGPT4: LLM generated concepts with manual refinement



Content-class association weights of Priv×ITM for PrivacyAlert: the ticker the line, the stronger the association

Takeaways

- LLMs/VLMs can be used for interpretable privacy predictions when employed for descriptive tasks
- PrivxITM achieves high performance without sacrificing interpretability
- Our method removes the need for human-annotated attributes for privacy classification

References

[1] Zhao et al., PrivacyAlert: a dataset for image privacy prediction, Int. AAAI Conf. on Web and Social Media, 2022

[2] Orekondy et al., Towards a visual privacy advisor: understanding and predicting privacy risks in images, ICCV, 2017

[3] Zhao et al., Deep gated multi-modal fusion for image privacy prediction, ACM Trans. Web, 2023

[4] Ayci et al., PEAK: explainable privacy assistant through automated knowledge extraction, arXiv, 2023