Cross-Modality Personalization for Retrieval (Supplementary Material)

Nils Murrugarra-Llerena   Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{nineil, kovashka}@cs.pitt.edu

As supplementary material, we provide quantitative results for our approach in the per task and joint setups for top-3 accuracy, rank and top-1 accuracy. We also show visualizations from our COCO data, the employed personality questionnaire and how personality affects interactions.

1. Combining content and style per task

We first show the benefit of combining content and style. Tables 1, 2, 3, 4, 5 and 6 show results for top-3 accuracy, rank and top-1 accuracy on six tasks on the *Ads* and *COCO* datasets, respectively, mapping each of our three modalities to each other modality. For each task, we show the performance of BASE, STYLE and OURS (combined). As discussed in the main text, the CONTENT method only makes sense in the case of retrieving gaze from captions, and vice versa. For the other tasks, we only use a subset of the constraints that content in general considers. We also show the performance of VEIT. We model all tasks separately i.e. we create three networks, one for each pair of modalities. Thus, the first/third, second/sixth, and fourth/fifth rows in each table correspond to the same network.

Our best result is for the rank measure (Tables 2 and 5), where our approach outperforms all other baselines in four out of the six tasks for both the *Ads* and *COCO* datasets. In this setup, our weakest result is for g2p/p2g, where VEIT outperforms our approach. We believe VEIT finds a latent link between these modalities, which allow easy retrieval in constrast to our method, as the latter does not use matrix factorization. Our best competitors for top-3 and top-1 accuracy are BASE and STYLE (Tables 1, 3, 4 and 6). However, overall from our comprised measures Tables 1 and 3 in our main text, our method performs strongest in the context of all metrics and all tasks. In contrast, other methods have inconsistent performance, i.e. they do well on some metrics but not others.

As our base network, we use VSE++ on Ads. Note that we also experimented with ADVISE from [4] as our base network, but it performed worse. ADVISE models image features, while we use a gaze-masked image. In particular,

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	0.2107	**0.2111**	N/A	0.206	0.2051
t2p	0.2625	**0.2894**	N/A	0.2806	0.2861
p2g	0.1671	**0.1754**	N/A	0.1643	0.1704
t2g	0.3783	0.4023	0.4384	0.2704	**0.4426**
g2t	0.3801	0.3745	0.4366	0.3074	**0.4463**
p2t	0.2556	0.2718	N/A	0.2741	**0.2768**
avg	0.2757	0.2874	0.1458	0.2505	**0.3046**

Table 1. Top-3 accuracy for the task-specific setup (higher is better) in the *Ads* dataset. N/A values were replaced with zero for average calculation.

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	**7.9912**	8.0199	N/A	8.0361	8.0718
t2p	7.3523	7.1445	N/A	7.0819	**7.0495**
p2g	**7.9241**	7.9949	N/A	8.0625	8.0259
t2g	5.6254	5.4213	5.1315	6.5926	**5.0393**
g2t	5.7305	5.7551	5.2292	6.6616	**5.1417**
p2t	7.4148	7.2403	N/A	7.1894	**7.1653**
avg	7.0064	6.9293	11.7268	7.2707	**6.7489**

Table 2. Rank for the task-specific setup (lower is better) in the *Ads* dataset. N/A values were replaced with fifteen (worst rank) for average calculation.

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	0.0838	**0.0829**	N/A	0.0769	0.0792
t2p	0.1213	0.1463	N/A	0.144	**0.15**
p2g	0.0398	0.0472	N/A	0.0431	**0.0495**
t2g	0.1088	0.119	0.1139	0.0764	**0.1241**
g2t	0.138	0.1514	0.1616	0.1157	**0.1648**
p2t	0.1121	0.1148	N/A	0.1218	**0.1264**
avg	0.1006	0.1103	0.0459	0.0963	**0.1157**

Table 3. Top-1 accuracy for the task-specific setup (higher is better) in the *Ads* dataset. N/A values were replaced with zero for average calculation.

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	0.2121	**0.2222**	N/A	0.2194	**0.2222**
t2p	0.2954	0.2926	N/A	**0.3102**	0.3074
p2g	0.1685	0.1556	N/A	**0.1759**	0.1639
t2g	0.4852	0.5371	0.6139	0.3269	**0.625**
g2t	0.4639	0.5204	0.5972	0.3657	**0.6065**
p2t	0.2722	0.2769	N/A	0.2787	**0.2833**
avg	0.3162	0.3341	0.2019	0.2795	**0.3681**

Table 4. Top-3 accuracy for the task-specific setup (higher is better) in the *COCO* dataset. N/A values were replaced with zero for average calculation.

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	**7.8537**	8.1509	N/A	8.0333	8.0917
t2p	7.0389	6.9482	N/A	7.0324	**6.8713**
p2g	**7.7972**	8.0685	N/A	8.0509	8.0407
t2g	4.7426	4.2713	3.7815	6.112	**3.6555**
g2t	4.8593	4.4833	3.8861	6.2833	**3.7352**
p2t	7.1241	6.9482	N/A	7.0306	**6.8695**
p2t	6.5693	6.4784	11.2779	7.0904	**6.2107**

Table 5. Rank for the task-specific setup (lower is better) in the *COCO* dataset. N/A values were replaced with fifteen (worst rank) for average calculation.

	Veit [3]	Base [1]	Content	Style	**Ours**
g2p	0.0982	**0.1074**	N/A	0.0972	0.1037
t2p	0.1361	0.15	N/A	0.1537	**0.1639**
p2g	0.0389	0.0454	N/A	**0.0481**	0.0463
t2g	0.1371	0.1593	0.1713	0.0805	**0.1945**
g2t	0.1954	0.2037	**0.2472**	0.1195	0.2463
p2t	0.1167	0.1185	N/A	0.1157	**0.1259**
avg	0.1204	0.1307	0.0698	0.1025	**0.1468**

Table 6. Top-1 accuracy for the task-specific setup (higher is better) in the *COCO* dataset. N/A values were replaced with zero for average calculation.

we masked the last convolution layer of Inception-v4 with our BubbleView gaze map. This procedure may hide some relevant information that ADVISE relies on. Also, ADVISE extracts regions of interest (ROI) from the image and finds an embedding space for the image and ROIs. However, in our approach, we do not employ the full image, instead, we use some salient locations, which could hamper the generated embedding space.

2. Joint modeling of all tasks

We present the top-3 accuracy, rank and top-1 accuracy for the joint setup with privileged information for the *Ads* data in Tables 7, 8 and 9. We exclude CONTENT because it does not apply to all modality pairs. Similarly to our summarized table from our main article, we outperform the baselines in three out of six tasks for top-3 accuracy (Table 7).

	Veit [3]	Base [1]	Style	**Ours**
g2p	0.2056	0.2042	**0.2083**	0.2051
t2p	0.2611	0.2852	0.3019	**0.3134**
p2g	0.1532	0.1787	**0.1815**	0.1792
t2g	0.3625	0.3843	0.2671	**0.4079**
g2t	0.382	0.3847	0.294	**0.412**
p2t	0.2528	0.2569	**0.2847**	0.281
avg	0.2695	0.2823	0.2563	**0.2998**

Table 7. Top-3 accuracy for joint setup (higher is better) for the *Ads* data.

	Veit [3]	Base [1]	Style	**Ours**
g2p	8.0843	8.0296	8.0236	**8.0111**
t2p	7.3778	7.2398	**6.8875**	6.9185
p2g	8.0676	8.012	**7.9732**	8.0093
t2g	5.6593	5.5903	6.7218	**5.3875**
g2t	5.6782	5.7245	6.7796	**5.4935**
p2t	7.394	7.3977	**7.0398**	7.0935
avg	7.0435	6.9990	7.2376	**6.8189**

Table 8. Rank for joint setup (lower is better) for the *Ads* data.

	Veit [3]	Base [1]	Style	**Ours**
g2p	0.0694	0.0754	**0.0778**	0.0768
t2p	0.1167	0.1426	0.1523	**0.1593**
p2g	0.0366	0.0403	**0.0472**	0.0435
t2g	0.0912	0.1102	0.0699	**0.1241**
g2t	0.1366	0.1449	0.1009	**0.1593**
p2t	0.1065	0.1116	0.1264	**0.131**
avg	0.0928	0.1042	0.0958	**0.1157**

Table 9. Top-1 accuracy for joint setup (higher is better) for the *Ads* data.

In terms of rank, STYLE outperforms our method in three tasks, t2p, p2g and p2t (Table 8), however, there is not a big difference with our method. Finally, as shown in Table 3 of our main text, our method is best overall across metrics and tasks.

3. Data visualization for COCO data

In Fig. 1, we show gaze and text samples from users on the same image. This is the complement to Fig. 2 from the main text, which shows samples on our Ads data. Each column shows results from the same two users; the top responses are from one and the bottom from another.

In the first column, we observe that the first user (in blue) is perhaps lazier than the second user (in red). Sentences from the first user are shorter and have fewer nouns than sentences from the second user. We also observe that the blue user explores a smaller part of the image, in contrast to the red user. From their personality responses, the second user is more conscious and neurotic, which is implies being a more analytical person.

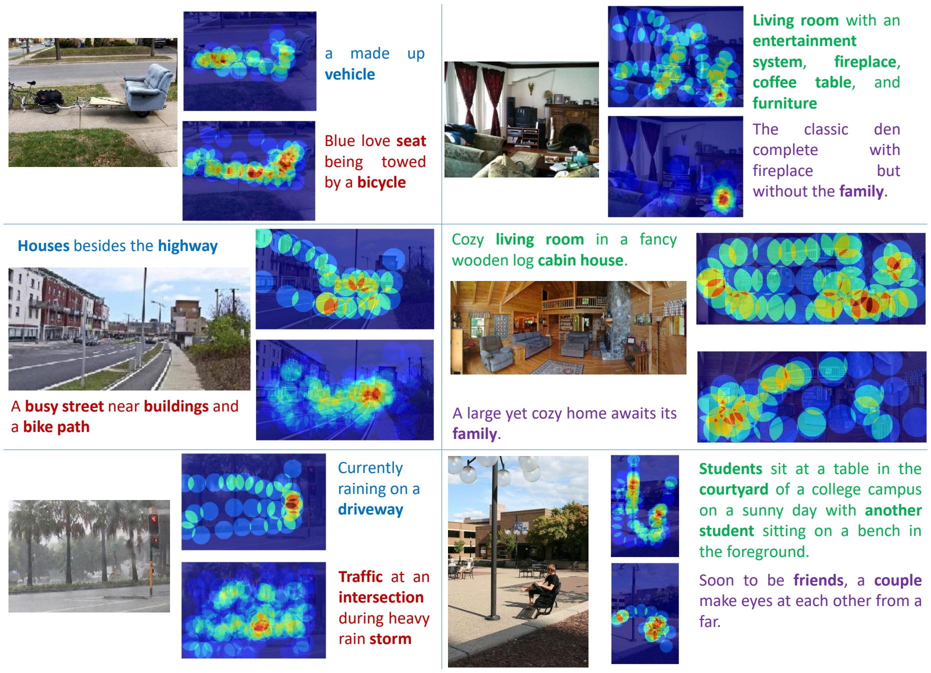

Figure 1. Text and gaze samples for different users on our *COCO* data. Each column shows three images annotated by two users.

	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
... is reserved	○	○	○	○	○
... is generally trusting	○	○	○	○	○
... tends to be lazy	○	○	○	○	○
... is relaxed, handles stress well	○	○	○	○	○
... has few artistic interests	○	○	○	○	○
... is outgoing, sociable	○	○	○	○	○
... tends to find fault with others	○	○	○	○	○
... does a thorough job	○	○	○	○	○
... gets nervous easily	○	○	○	○	○
... has an active imagination	○	○	○	○	○

Table 10. Personality survey [2] as shown to Amazon Mechanical Turkers. Each question starts with "I see myself as someone who..."

In the second column, we observe that the first user (in green) is more analytical than the second user (in purple), who happens to be a more empathic person. For example, the first user annotates the image with objects present in the image, and the second viewer emphasizes relationships with others (i.e. family, friends, couple, etc). From the personality responses, the second person is more agreeable than the first one. Agreeableness is closely related to generosity, emphasis, and sympathy, which relates to making connections with others.

4. Personality questionnaire

The complete personallity survey [2] is shown in Table 10. This survey measures five dimensions of personality: neuroticism, extraversion, openness, aggreableness and conscientiousness. Each question queries for a response in the range from "disagree strongly" to "agree strongly". Neuroticism is closely related to people tendencies for anxiety, hostility, depresion and low self-steem, while extraversion for positive, energetic and encouraging tendencies. Openness encompass personality traits such as curiosity, artistry, flexibility and wisdom, while aggreableness is related to kindness, generosity, empathy, altruism and trusting others. Finally, conscientiousness measures people traits such as efficiency, reliableness and rationality.

5. How personality affects interactions

We show some correlations that help explain how personality determines the text a user produces. We isolated responses to each personality question, and retrieved the text matching the positive (agreeging with question) and negative users. Table 11 shows frequently employed words: e.g. *less reserved* people may use *like*, *love* because they express their feelings more, and *more relaxed* people may use stronger positive adjectives.

reserved	POS - [beautiful, care, fun, mini, oreos] NEG - [beer, **like**, look, **love**, makeup]
relaxed	POS - [better, **fun, great**, help, need] NEG - [animals, beautiful, legos, like, look]

Table 11. The five most frequent words for personality types. Words shared by pos/neg group are removed.

References

[1] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. In *British Machine Vision Conference (BMVC)*. Springer, 2018.

[2] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality (JRP)*, 2007.

[3] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[4] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *European Conference on Computer Vision (ECCV)*. Springer, 2018.