Personalized Image Descriptions from Attention Sequences

Supplementary Material

1. Overview

The supplementary material is organized as:

- Supplementary evaluations of DEPER: (1) additional personalized image captioning baselines Sec. 2.1, (2) quantitative analysis of DEPER outputs Sec. 2.2, (3) a user study assesses both personalization and generalization to new attention/description styles Sec. 2.3.
- Supplementary analysis: (1) Evaluation of model-generated (simulated) attention supervision when human attention data is unavailable Sec. 3, (2) Analysis of the prompt-tuning design for generalization Sec. 4.
- Ablation study on (1) the influence of the number of samples used to construct subject embeddings Sec. 5, and (2) further ablation on unseen subjects Sec. 5.2.
- Further implementation details, including (1) network architecture Sec. 6.1, (2) prompt design Sec. 6.2, and (3) metric computation algorithms Sec. 6.3.
- Additional qualitative examples of generated image descriptions Sec. 7.

2. Supplementary evaluations of DEPER

In this section, we evaluate our method on the standard personalized captioning benchmark YFCC100M [12] with more baselines in Sec. 2.1, and present quantitative analyses of DEPER's outputs in Sec. 2.2.

2.1. Results on YFCC100M

YFCC100M [15] is a 100-million–item Flickr dataset of user-uploaded images and videos (with titles, descriptions, and tags), where titles/descriptions are treated as captions. The preprocessed subset from CSMN [12] is widely used in personalized image captioning models to test their performance, which contains 46M captions from 6197 users in the training set, and 255 subjects in the testing set. We compare against CSMN [12] and Wang *et al*. [17], two personalized image captioning models that use this publicly available train-test split, ensuring a fair comparison. We achieve SOTA performance on this dataset versus both prior approaches.

Table 1. **Image captioning performance on YFCC100M**[12, 15]. CSMN[12] and Wang *et al*. [17] are two baselines using the public dataset split.

Models	B1	B4	M	R	C
CSMN[12]	0.116	0.003	0.036	0.060	0.111
Wang *et al*. [17]	0.100	0.008	0.064	0.097	0.170
Ours	**0.134**	**0.019**	**0.084**	**0.142**	**0.211**

2.2. Quantitative Results on DEPER's output

This section reports the quantitative results corresponding to DEPER's qualitative examples in Figure 4 in the main paper. DEPER outputs a predicted subject ID and a reconstructed attention trajectory. In Tab. 2, we present the F1 score for subject classification on seen and unseen splits, where the former uses DEPER's trained classifier and the latter uses KNN over DEPER's embeddings. We also measure the intersection rate between predicted and ground-truth bounding-box sequences, counting a box as a hit if it overlaps with the ground-truth box at the same or adjacent time step. This measures both the spatial and temporal reconstruction performance. The quantitative results demonstrate that DEPER effectively distinguishes individual subjects and maintains strong performance on unseen ones, showing that its learned subject embeddings generalize beyond the training population. On the challenging Kollenda *et al*. [8] dataset – where all 30 subjects view the same images – DEPER still achieves solid recognition, indicating its ability to disentangle personality traits from image content and capture the underlying factors that drive human viewing and description behaviors.

Table 2. **Quantitative results of DEPER's output**. F1 is the F1 score of subject ID classification. Intersec. is the intersection rate of the predicted sequence of bounding boxes and ground-truth. He *et al*. [4] does not have unseen split due to the limited number of subjects.

Datasets	COCO-LN[13]	Flickr30k-LN[13]	Kollenda *et al*. [8]	He *et al*. [4]
Seen F1	0.840	0.963	0.549	0.698
Unseen F1	0.644	0.851	0.310	–
Seen Intersec.	0.411	0.351	0.408	0.352
Unseen Intersec.	0.398	0.316	0.361	–

2.3. User Study

We evaluate our method through a user study consisting of two stages. (i) Data collection: Each user describes **5 COCO-train images** while we record mouse trajectories (users point to what they describe, same pipeline as COCO-LN), from which DEPER infers a user embedding to conditions Qwen2-VL to caption **15 validation images**. (ii) User preference: Users compare Qwen-PT and Qwen-DEPER captions and select the closer match to their style and applicable reasons. We report WIN rates over **8 users** in Tab. 3, along with selected reason and caption metrics. Reason r (%) is $\frac{\#\text{r selected}}{\#\text{wins}}$ for that method. The user study interfaces in steps (i) to collect description and attention trajectories are shown in Fig. 1.

We also present qualitative results in Fig. 2, comparing

our method with Qwen-PT, which further demonstrates its generalization ability.

Table 3. User study results.

Methods	WIN	Detail	Coverage	Order	Phrasing	ROUGE	CIDEr
Qwen-PT	22%	61%	81%	23%	9%	0.199	0.188
Qwen-DEPER	78%	67%	73%	48%	29%	0.284	0.302

3. Alternative Training Approaches with Simulated Attention Supervision

Although collecting new image–description data with mouse tracking is straightforward, most existing caption datasets lack human attention labels. To test whether our method still applies in this setting, we run an additional experiment on Flickr30k-LN [13] where we replace ground-truth attention trajectories with simulated ones. Even under this weaker supervision, we still observe clear gains over using purely linguistic subject embeddings from DEPER shown in Tab. 4.

To generate simulated attention trajectories, we use GroundingDINO [11] to localize each noun with a bounding box and concatenate these boxes in noun order to form model-generated attention trajectories. These trajectories are beneficial because they preserve the key components of human attention for image descriptions: which regions are attended, in what order, and with what level of detail, while also encoding basic object–object relations that continuously guide the description. However, the marginal gap between model-generated and human-annotated performance is due to the precise localization signal provided by ground-truth human attention trajectories.

Table 4. **Model-generated attention supervision**. To enable DEPER on datasets without human attention annotations, we compare three variants: (i) w/o Traj., which uses only descriptions to learn subject embeddings; (ii) Sim. Traj., which replaces human trajectories with model-generated (simulated) trajectories; and (iii) GT Traj., our full model with ground-truth human attention. Results are reported on Flickr30k-LN[13].

Traj. Type	B4	M	R	C	P	OSS	CLS
w/o Traj.	0.222	0.247	0.500	0.770	0.659	0.379	0.649
Sim. Traj.	0.272	0.264	0.512	0.775	0.665	0.397	0.768
GT Traj.	0.312	0.272	0.518	0.789	0.671	0.408	0.796

4. Adaptability Analysis

A core goal of DEPER is to learn subject embeddings that remain portable and generalize to new individuals rather than overfitting to a fixed set of training users. To do so, we employ prompt tuning, where each subject is represented only through the explicit embedding z_s inserted as a token in the VLM's input space. The backbone remains frozen, and no subject-specific updates are written into its internal parameters. This design forces all personalization to flow through $\mathbf{z}_s$, keeping the encoding of subject-level traits compact and preventing identity information from leaking into the model weights.

Table 5. Comparison of DEPER with different fine-tuning strategies of Qwen[16]: LoRA[6] vs. PT(Prompt-tuning)[9] on seen and unseen splits. The results highlight a trade-off between seen and unseen subjects: LoRA performs better on seen data but degrades substantially on unseen data. Results are reported on Flickr30k-LN[13].

Split	Method	B4	M	R	C	P	OSS	CLS
Seen	DEPER+LoRA	**0.387**	**0.310**	**0.569**	**1.077**	**0.691**	**0.460**	**0.818**
	DEPER+PT	0.312	0.272	0.518	0.789	0.671	0.408	0.796
Unseen	DEPER+LoRA	0.191	0.216	0.404	0.369	0.571	0.301	0.479
	DEPER+PT	**0.202**	**0.232**	**0.410**	**0.382**	**0.610**	**0.329**	**0.625**

This differs from approaches such as LoRA or full-model fine-tuning, which update large portions of the model and therefore tend to specialize to the subjects seen during training. While this can improve performance on those subjects, it usually offers weaker transfer because part of the personalization behavior is tied to adapted parameters that do not generalize to new users, as can be seen from Tab. 5.

Further, we assess this behavior with a swap test on unseen users. In this test, we find that replacing the correct z_s with an embedding from another subject leads to a larger performance drop under prompt tuning than under LoRA: a 30.9% drop for DEPER+PT versus 19.5% for DEPER+LoRA. To obtain these numbers, for each test pair (I, D_s, s) we keep I and D_s fixed and substitute the correct z_s with $z_{s'}$ from a different subject; if personalization depends on z_s, this mismatch should degrade performance. The larger drop under prompt tuning indicates a stronger reliance on the explicit subject embedding, whereas LoRA retains more subject-specific information in its adapted parameters.

5. Supplementary Ablation

5.1. Selection of K

Table 6. **Ablation on the number of samples for subject embedding computation**. # K denotes the number of image-description-trajectory triplets per subject in the seen split used to compute the subject embedding by averaging their DEPER output. Results indicate the consistent subject embedding across different image contents. Results are reported on the Flickr30k-LN[13].

# K	B4	M	R	C	P	OSS	CLS
10	0.310	0.268	0.517	0.792	0.668	0.407	0.793
50	0.312	0.270	0.518	0.794	0.670	0.407	0.792
100	0.312	0.272	0.518	0.789	0.671	0.408	0.796

Table 7. Few-shot ablation on Flickr30k-LN. Higher is better.

Metric	Text-only	Pseudo-traj	16 tokens	Offset&Cut	Ours
ROUGE	0.387	0.401	0.399	0.409	**0.410**
CIDEr	0.365	0.374	0.379	0.379	**0.382**

We study the effect of varying K, the number of samples used to construct subject embeddings on the seen set. As shown in Tab. 6, changing K has little impact on description quality, suggesting that DEPER effectively disentangles subject personality from image content.

5.2. Ablation on unseen split

To examine key factors affecting practical deployment, we ablate three settings: (1) training DEPER using text only, without attention supervision or inputs (text-only); (2) using simulated attention trajectories for unseen users (pseudo-traj; see Sec. 3 for details); and (3) introducing noisy attention trajectories by randomly dropping 10% of fixations and adding small offsets to the remaining fixations (Offset&Cut). As shown in Tab. 7, DEPER remains stable across these settings, demonstrating robustness and flexibility in practical scenarios.

We further study the effect of the number of tokens used to represent a user in Tab. 7. Specifically, we replace the single token with 16 tokens in both DEPER's output and the prompt of Qwen2-VL, following the practice in LLaVA [10] of replacing image token IDs with image embeddings. Results show that a single token suffices under our formulation. This is because the subject embedding is not intended to encode fine visual details, but to capture consistent, largely image-invariant user preferences in the caption domain. Image content is provided by Qwen's world knowledge, while the embedding modulates personalization, making one token sufficient.

5.3. Ablation on hyperparameters

We conduct an ablation study on the scaling factor λ, which controls the contribution of the contrastive loss across the two training stages. As shown in Tab. 8, $\lambda = 0.1$ achieves the best performance. This is largely because it balances the magnitude of the contrastive loss with other objectives, preventing any single loss term from dominating the optimization process.

Table 8. Ablation on the hyperparameter λ.

λ	B4	M	R	C	P	OSS	CLS
1.0	0.309	0.264	0.515	0.778	0.665	0.406	0.789
0.5	0.310	0.268	0.517	0.785	0.670	0.407	0.791
0.1	0.312	0.272	0.518	0.789	0.671	0.408	0.796

6. Implementation Details

6.1. Network Design

Trajectory Preprocessing. We tailor to the specific source (gaze fixations or mouse movements) to get the attention trajectory $T_s = \{(\mathbf{b}_i, \tau_i)\}_{i=1}^{M}$, where each $\mathbf{b}_i$ is one of M bounding boxes with a duration of τ_i.

Specifically, for mouse movement (COCO-LN and Flickr30k-LN) [13], we follow the method of [18]: (1) extract nouns from the descriptions using spaCy [5], (2) for each noun, we gather all mouse position samples that temporally overlapped the noun's utterance, and (3) we create the minimal rectangle to cover all mouse movement, forming one bounding box per noun.

For eye-tracking datasets [4, 8], we use SAM2 [14] to segment an object mask at each fixation, convert it into a bounding box, and apply non-maximum suppression to retain precise gaze-related objects. Duration of each noun is accumulated from all mouse movements/fixations that belong to this noun.

These object-level bounding boxes and their durations indicate *which* objects are likely to be mentioned and *how* much detail they receive. They also encode spatial and ordering regularities (e.g., nearby objects are often described consecutively) that cannot be recovered from the descriptions alone.

Input Encoders. We use three encoders to obtain modality-specific features: the image features $\mathbf{V} \in \mathbb{R}^{N_I \times D}$, the description features $\mathbf{L} \in \mathbb{R}^{N_C \times D}$, and the trajectory features $\mathbf{Tr} \in \mathbb{R}^{N_M \times D}$, where N_I, N_C, and N_M denote the numbers of image patches, text tokens, and bounding boxes, respectively. Specifically, the trajectory features are obtained by mean-pooling the image patch tokens that intersect with each bounding box.

Contrastive Loss. We adopt a supervised contrastive loss following SupCon [7]. Instead of computing the loss only within the current mini-batch, we maintain a MoCo-style memory bank [3] that stores subject embeddings from previous batches. This is important in our setting because the batch size is much smaller than the number of subjects, so a single mini-batch may contain no positive pairs for some subjects, and each embedding would otherwise see only a very limited set of negatives. The memory bank holds up to 4096 samples and is updated with a FIFO policy. The exact loss computation is summarized in Algorithm 1.

6.2. Prompt Design

In this section we show the two different prompts for Qwen+DEPER (ours) and the Qwen few-shot baseline.

Prompt for Qwen+DEPER. Provided at inference for all DEPER evaluation, including seen and unseen subjects.

```
System:  You are a personalized
```

Algorithm 1 Supervised Contrastive Loss [7] with Memory Bank [3]

1: **Input:** Batch embeddings and subject IDs $\{(\mathbf{z}_i, y_i)\}_{i=1}^{A}$, memory bank of previous embeddings $\{(\mathbf{v}_k, \tilde{y}_k)\}_{k=1}^{N}$, temperature τ

2: L2-normalize all embeddings:

$$\hat{\mathbf{z}}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2, \quad i = 1, \ldots, A$$
$$\hat{\mathbf{v}}_k = \mathbf{v}_k / \|\mathbf{v}_k\|_2, \quad k = 1, \ldots, N$$

3: For each anchor i, compute similarity to all bank entries:

$$s_{i,k} = \frac{\hat{\mathbf{z}}_i^\top \hat{\mathbf{v}}_k}{\tau}, \quad k = 1, \ldots, N$$

4: Convert to a softmax distribution over the bank:

$$p_{i,k} = \frac{\exp(s_{i,k})}{\sum_{m=1}^{N} \exp(s_{i,m})}, \quad k = 1, \ldots, N$$

5: Define positives for anchor i as all bank entries with the same subject ID:

$$P(i) = \{\, k \mid \tilde{y}_k = y_i \,\}, \quad |P(i)| \geq 1$$

6: Compute per-anchor supervised contrastive loss:

$$\ell_i = -\frac{1}{|P(i)|} \sum_{k \in P(i)} \log p_{i,k}$$

7: **Output:** Batch-averaged contrastive loss

$$\mathcal{L}_{\text{con}} = \frac{1}{A} \sum_{i=1}^{A} \ell_i.$$

```
vision-language model.
User: [IMAGE] <image> Write a detailed
description for this photo in the style of
{<subj>} .
```

Prompt for Qwen few-shot baseline. Provided at inference to test the few-shot ability of Qwen [16] in personalized image description generation.

```
System: You are a personalized
vision-language model. You are given
several support examples that demonstrate
how this subject describes images. Infer
this subject's style and use it to caption
the test image.
User: SUPPORT EXAMPLES
[IMAGE_1] Caption: {CAPTION_1} Trajectory:
{BBOXES_1}
[IMAGE_2] Caption: {CAPTION_2} Trajectory:
{BBOXES_2}
...
```

Algorithm 2 Object Sequence Score (OSS)

1: **Input:** Prediction caption $\hat{y}$, reference caption y, gap penalty γ

2: Use spaCy to extract NOUN/PROPN tokens $\text{Nouns}(y)$ and $\text{Nouns}(\hat{y})$:
$P = (p_1, \ldots, p_N) = \text{Nouns}(\hat{y})$, $R = (r_1, \ldots, r_M) = \text{Nouns}(y)$

3: Define token-level match score $s(p_i, r_j)$ as

$$s(p_i, r_j) = \begin{cases} 1.0, & \text{if } p_i = r_j \text{ (exact match)} \\ 0.8, & \text{if stems}(p_i) = \text{stems}(r_j) \\ 0.7, & \text{if } p_i \text{ and } r_j \text{ are synonyms} \\ 0, & \text{otherwise.} \end{cases}$$

4: Compute an alignment score $A(P, R)$ using the standard Needleman–Wunsch algorithm on sequences P and R with match function $s(\cdot, \cdot)$ and gap penalty γ.

5: **Output:** Normalize by max sequence length:

$$\text{OSS}(\hat{y}, y) = \frac{A(P, R)}{\max(N, M)} \in [0, 1].$$

```
User: TEST IMAGE[TEST_IMAGE] <image>
Describe this image in the same style as
the
support examples.
```

6.3. Metrics Computation

Object Sequence Score (OSS). Our proposed object sequence score metric generalizes the sequence score and semantic sequence score metrics which are commonly used to evaluate dynamic attention prediction models [1, 2, 19]. The detailed implementation is shown in Algorithm 2.

Description Classficiation Accuracy (CLS). The detailed implementation of CLS is shown in Algorithm 3.

Other Metrics. The BLEU, METEOR, ROUGE, and CIDEr used in main paper are typically computed by comparing one prediction to multiple references. We slightly modify them to enforce one-to-one prediction–reference matching. This better evaluates how well a prediction of a subject matches the ground truth of that subject.

7. Qualitative Results of Description Generation

We show more qualitative results in Fig. 3. The results show that our model captures subject-specific diversity in descriptions, such as which objects are mentioned, the order of objects, and the details. It also demonstrates strong generalization on Kollenda *et al.*'s dataset by producing human-aligned captions for both seen and unseen subjects on the

Algorithm 3 Subject Classification Accuracy with COCO Metrics

Input: For each image i, a set of m_i description pairs $\{(\hat{y}_i^j, y_i^j)\}_{j=1}^{m_i}$, where $\hat{y}_i^j$ is the prediction and y_i^j is the ground-truth caption for subject j; similarity metric M (BLEU-4, METEOR, ROUGE-L, CIDEr).
Output: Overall subject-classification accuracy.
Initialize total hits $H \leftarrow 0$, total pairs $T \leftarrow 0$.
for each image i **do**
    for $j = 1$ to m_i **do**

$$s_{j,k} = M(\hat{y}_i^j, y_i^k), \quad k = 1, \ldots, m_i.$$

        Let $k^* = \arg\max_k s_{j,k}$.
        if $k^* = j$ **then** ▷ paired reference has highest score
            $H \leftarrow H + 1$
        end if
        $T \leftarrow T + 1$
    end for
end for
return CLS $= \dfrac{H}{T}$.
Repeat for different M and average.

same image.

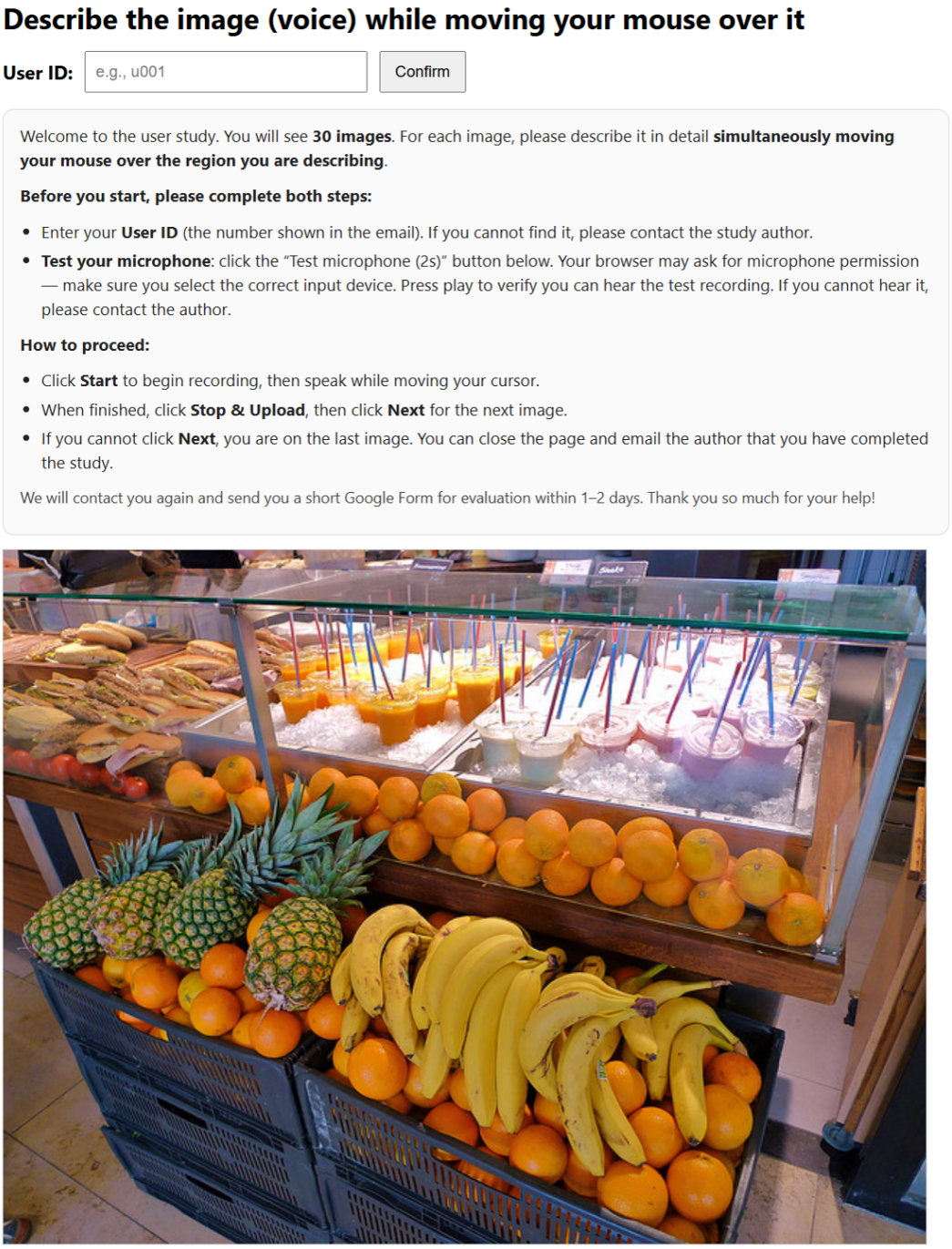


Figure 1. User study interface for collecting the support set and the test set of new users.

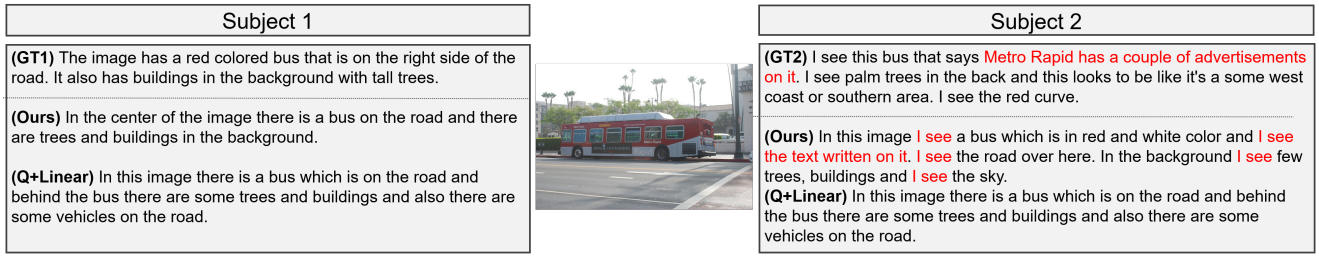


Figure 2. **Qualitative Results** on unseen subjects from the user study. Despite differences in linguistic style and content compared to the training set (COCO-LN), DEPER successfully captures distinct personality traits from a small support set.

Seen

Prompt: Write a detailed description of this photo in the style of **<axke_1>:**

(Ours) A bear is sitting on a rock. There are some plants behind it.

(GT) A bear cub is standing on a stone beside a bear. There are trees around them.

Prompt: Write a detailed description of this photo in the style of **<wons_2>:**

(Ours) In this image I can see a bear which is black in color is standing on the rock and I can see another bear which is black in color is standing on the rock. In the background I can see the trees and the sky.

(GT) In this image I can see two animals which are black in color and i can see rocks and few trees.

Unseen

Prompt: Write a detailed description of this photo in the style of **<oans_3>:**

(Ours) Two men are jumping in the air, they are holding a ball in their hands. In the background there is a wall, door.

(GT) There are two men playing football and we can see both are jumping. In the background there is a wall, door.

Prompt: Write a detailed description of this photo in the style of **<xcmq_4>:**

(Ours) Two persons are playing a game in the court. In the background there is a wall and a fire extinguisher.

(GT) Two persons are playing basketball. And this is a basketball court. In the background there is wall with doors, and a fire extinguisher.

Seen

Prompt: Write a detailed description of this photo in the style of **<svri_1>:**

(Ours) In this image, we can see a dog and there is a plastic bag with a cartoon on it. In the background, there is a wall.

(GT) In this image, we can see an animal on the carry bag and we can see pictures on the carry bag. There are walls. At the bottom, there is a flooring.

Prompt: Write a detailed description of this photo in the style of **<ascj_2>:**

(Ours) This image consists of a white and black dog and a plastic bag. In which, we can see a bottle. At the bottom, there is a floor. In the background, there is a wall.

(GT) This image consists of a dog in white and black color is playing with a cover. At the bottom, there is floor. In the background, we can see a wall.

Unseen

Prompt: Write a detailed description of this photo in the style of **<nciw_3>:**

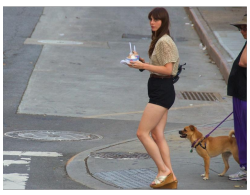
(Ours) In the image I can see a person wearing a life. I can see the water and a few things in the water.

(GT) In this image, I can see a girl wearing swimsuit is swimming in the water. In which I can see some objects in the top right corner.

Prompt: Write a detailed description of this photo in the style of **<owlk_4>:**

(Ours) An outdoor image. In the water there is a person wearing a life jacket. In the water there is a person wearing a costume. In the water there is a object. At the top of the image there is water.

(GT) This image is taken outdoors. In this image a kid is swimming in the water. The kid has worn a floating jacket. At the top right of the image there is an object in the water.



Seen

Prompt: Write a detailed description of this photo in the style of **<qujc_1>:**

(Ours) A woman stands on the sidewalk and holds a ice cream in her hand. In the background there is a dog.

(GT) A woman with two ice cream runs across the street. Behind her is a woman with a dog.

Prompt: Write a detailed description of this photo in the style of **<asin_2>:**

(Ours) Woman with a short skirt and a belt, she held a ice cream in her hand and a dog and a woman were behind her.

(GT) A woman who goes over a street wears relative short clothes and has a ice cream maker in her hand. Behind her was a woman who goes for a walk with her dog.

Unseen

Prompt: Write a detailed description of this photo in the style of **<cykj_3>:**

(Ours) A woman walks down the street and a dog is on a leash next to her.

(GT) One women go over a crosswalk and one woman has a dog on a leash.

Prompt: Write a detailed description of this photo in the style of **<xcyw_4>:**

(Ours) A woman is on the street and holds two ice creams in her hand.

(GT) A woman wants to cross the street and carries two ice cream cups in her hand.

Figure 3. Qualitative results across datasets on seen and unseen splits. Text differences between two subjects are highlighted in red. From top to bottom: COCO-LN [13], Flickr30k-LN [13], and Kollenda *et al*. [8].

References

[1] Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16206–16216, 2025. 4

[2] Souradeep Chakraborty, Ruoyu Xue, Rajarsi Gupta, Oksana Yaskiv, Constantin Friedman, Natallia Sheuka, Dana Perez, Paul Friedman, Won-Tak Choi, Waqas Mahmud, et al. Measuring and predicting where and when pathologists focus their visual attention while grading whole slide images of cancer. *Medical Image Analysis*, page 103752, 2025. 4

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3, 4

[4] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8529–8538, 2019. 1, 3

[5] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3, 4

[8] Diana Kollenda, Anna-Sophia Reher, and Benjamin de Haas. Individual gaze predicts individual scene descriptions. *Scientific Reports*, 15(1):9443, 2025. 1, 3, 8

[9] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2

[12] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Towards personalized image captioning via multimodal memory networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):999–1012, 2018. 1

[13] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 1, 2, 3, 8

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[15] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1

[16] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2, 4

[17] Xuan Wang, Guanhong Wang, Wenhao Chai, Jiayu Zhou, and Gaoang Wang. User-aware prefix-tuning is a good learner for personalized image captioning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 384–395. Springer, 2023. 1

[18] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024. 3

[19] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024. 4