# EFSA: Episodic Few-Shot Adaptation for Text-to-Image Retrieval

## Supplementary Material

In the following sections, we present additional results and a more extensive qualitative evaluation.

## 7. Experiments with SigLIP

In Table 6, we provide multi-domain results with a more recent and performant vision-language model, SigLIP (ViT-SO400M-14) [41]. The strength of SigLIP over CLIP is evident in the Recall@k scores for the zero-shot baseline, all over 10 points higher for SigLIP compared to CLIP (see Table 3). Even with this stronger backbone, EFSA proves effective, yielding highest Recall@1 scores on 7 out of 8 datasets, the odd one out being yet again the Books dataset. That being said, the improvement is less pronounced here compared to the CLIP setting: the average Recall@1 increases from 34.48 to 36.15.We hypothesize that the smaller performance gain is attributable to the stronger and more robust SigLIP backbone, which is inherently better at handling hard negatives.

## 8. Effect of Caption Generation Prompts

Figure 6 shows how retrieval performance changes with different prompts for generating image captions. We tested prompts that varied in length constraints, from no word limit to a maximum of 10, 20, 30, or 40 words. Overall, the choice of prompt has less than 1 point impact across Flickr30k and ArtCap. Performance improves when captions increase from 10 to 20 words but starts to decline as the word count goes beyond 20.

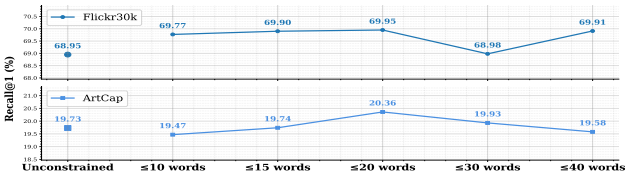



Figure 6. Effects of various caption generation prompts.

## 9. Qualitative Analysis

Figure 7 presents a qualitative comparison between zero-shot CLIP and EFSA in terms of the top-4 retrieved images on the ArtCap and TextCap datasets, with the synthetic captions for the images also included. We observe that the synthetic captions exhibit considerable semantic overlap with the query text. Notably, LLaVA accurately interprets text present within the images and incorporates it into the captions. Yet, as discussed in §4.2, a simple text-to-text retrieval approach does not prove effective here. EFSA instead enables the backbone model to learn from the image-caption pairs, leveraging not only information from the ground-truth image but also from the hard negatives surrounding it. Using this information to build more accurate representations for the images in the retrieval pool, the EFSA-modified CLIP can correclty re-rank the ground-truth image to the top position.

## 10. Storage Overhead Analysis

We compare the storage requirements introduced by EFSA—due to storing additional top-$k$ captions—with the zero-shot baseline that stores only image embeddings.

**Baseline Storage:** We estimate the storage requirement of the zero-shot T2I baseline as:

$$\texttt{Total Storage} = \langle \textit{size of retrieval pool} \rangle \times \langle \textit{bytes per cached image rep} \rangle. \tag{3}$$

Each cached image embedding is a 768-dimensional `float32` vector (as used in CLIP-base), requiring:

$$768 \times 4 = 3072 \text{ bytes per image.} \tag{4}$$

**EFSA Caption Overhead:** EFSA adds one generated caption per image, stored as a sequence of CLIP vocabulary token IDs, each encoded as a `uint16`. Assuming an average of 30 tokens per caption, the additional storage required is:

$$30 \text{ tokens} \times 2 \text{ bytes/token} = 60 \text{ bytes per image.} \tag{5}$$

**Relative Overhead:** The relative increase in storage per image is given by:

$$\frac{60}{3072} \approx 0.0195 \ (\approx 2\%). \tag{6}$$

Thus, EFSA introduces a minimal storage overhead of approximately 2% compared to the zero-shot baseline, while providing measurable gains in both open- and closed-domain retrieval performance.

## 11. Effect of Loss Function

Table 7 shows the impact of different training objectives on retrieval performance across the COCO and ArtCap

Table 6. Text-to-image retrieval performance in a multi-domain setting **with a SigLIP backbone**. Results are reported for Zero-Shot (Z.S), Fine-Tuning (F.T), Text-to-Text (T2T), and Episodic Few-Shot Adaptation (EFSA). The results demonstrate that EFSA consistently surpasses other methodologies, particularly on Recall@1 in this complex retrieval setup.

| **Multi-domain** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **COCO** | | | **Flickr30k** | | | **Books** | | | **NASA** | | | | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | | | |
| Z.S | 39.1 | 61.71 | 70.31 | 49.23 | 72.15 | 79.29 | **32.27** | **49.98** | **54.9** | 14.9 | 27.22 | **35.66** | | | |
| F.T | 30.23 | 53.32 | 63.21 | 34.74 | 58.89 | 67.72 | 7.72 | 16.3 | 20.89 | 3.61 | 10.12 | 13.73 | | | |
| T2T | 18.27 | 32.50 | 39.27 | 20.44 | 33.00 | 39.46 | 0.98 | 2.05 | 2.75 | 2.16 | 4.81 | 6.26 | | | |
| EFSA | **42.61** | **64.69** | **72.27** | **52.49** | **75.08** | **80.74** | 31.55 | 48.44 | 53.84 | **15.18** | **27.46** | 34.69 | | | |
| | **VizWiz** | | | **TextCap** | | | **ArtCap** | | | **SciCap** | | | **Average** | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| Z.S | 31.99 | 49.24 | 55.55 | 58.75 | 73.00 | 77.93 | 13.21 | 28.67 | 36.89 | 36.46 | **50.49** | **56.53** | 34.48 | 51.55 | 58.38 |
| F.T | 22.98 | 39.43 | 46.54 | 43.74 | 60.08 | 66.34 | 10.65 | 24.03 | 31.87 | 5.96 | 12.33 | 15.56 | 19.95 | 34.31 | 40.73 |
| T2T | 14.07 | 24.28 | 28.90 | 22.68 | 33.32 | 38.40 | 5.33 | 12.27 | 16.41 | 8.63 | 13.79 | 16.69 | 11.57 | 19.50 | 23.51 |
| EFSA | **33.66** | **50.85** | **56.28** | **60.95** | **74.52** | **78.94** | **15.45** | **31.43** | **38.52** | **37.36** | 50.33 | 55.4 | **36.15** | **52.85** | **58.83** |

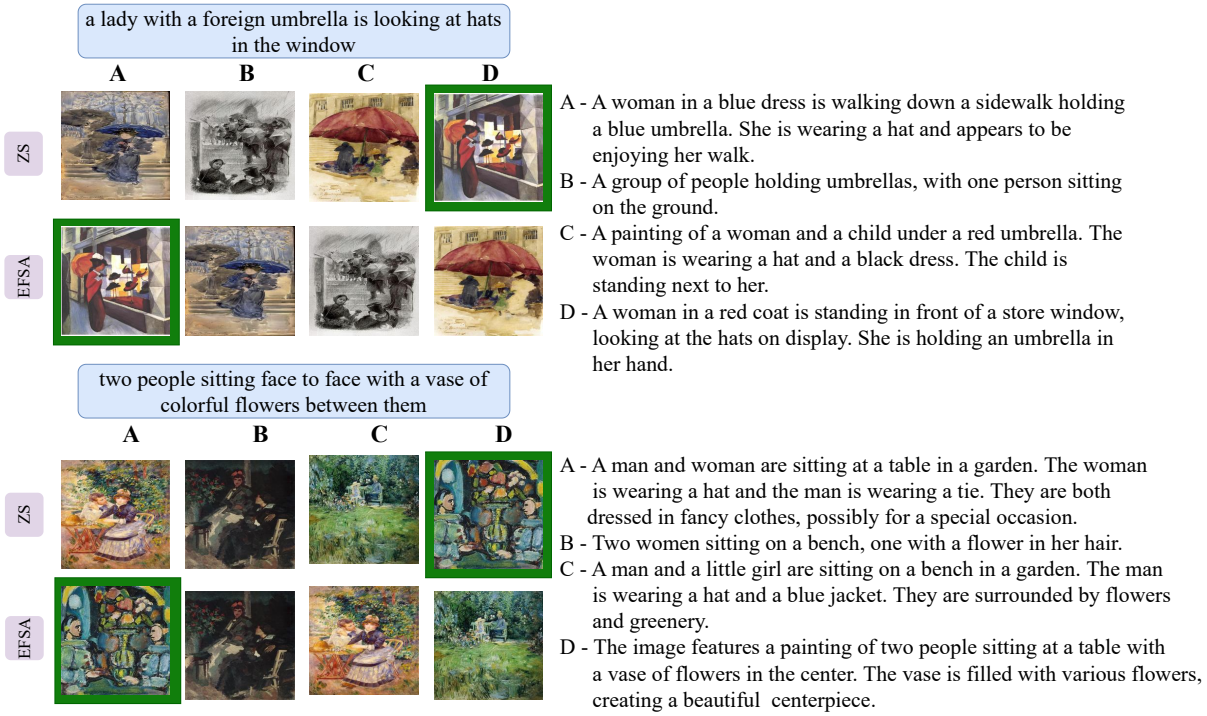
Table 7. Effect of various loss Functions on text-to-image Retrieval performance. A weighted combination of contrastive and hinge loss enhances retrieval performance.

| **Loss Function** | **COCO** | | | **ArtCap** | | |
|---|---|---|---|---|---|---|
| | **R@1** | **R@5** | **R@10** | **R@1** | **R@5** | **R@10** |
| Hinge | **30.15** | 50.78 | 57.79 | 10.91 | 22.73 | 27.47 |
| Contrastive | 28.23 | 49.20 | 56.47 | 10.33 | 21.66 | 26.61 |
| Combined | 30.14 | **50.96** | **57.82** | **11.13** | **22.83** | **27.50** |

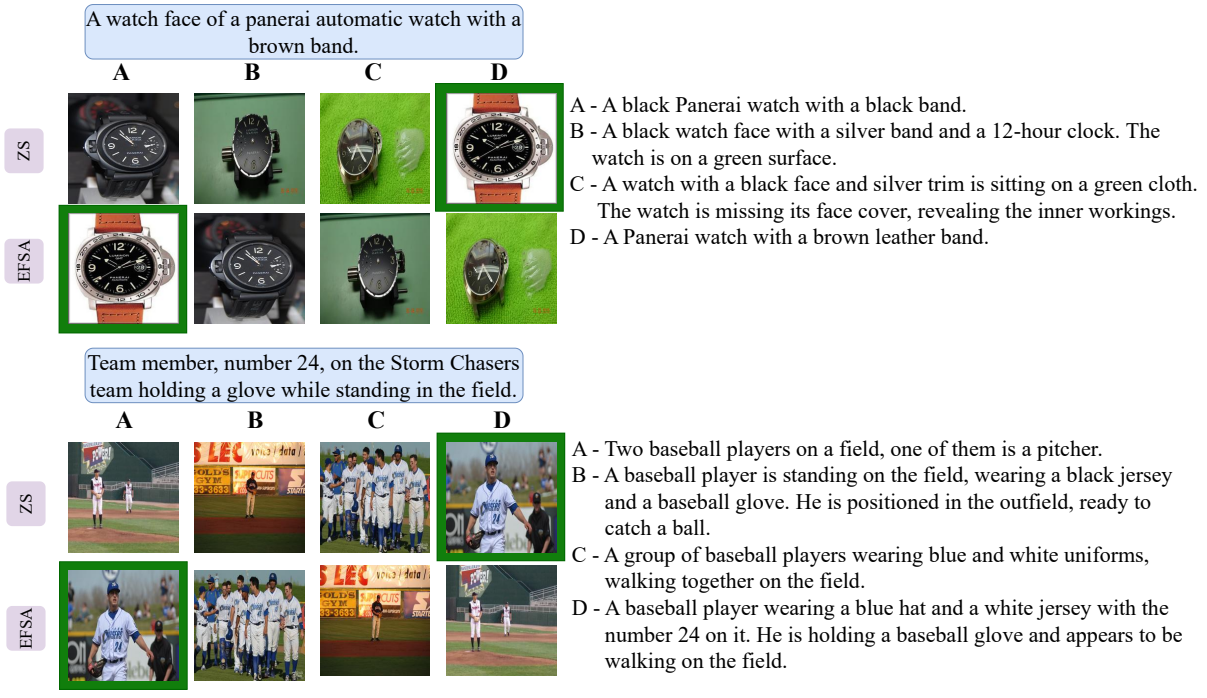
datasets. Hinge loss proves more effective than the contrastive loss here, likely due to the nature of the data, consisting of highly-similar images with highly similar captions. Through the margin in the hinge loss, these data points are being actively pushed away from each other, forcing the model to pay attention to subtle differences. Although the contrastive loss is less performant than the hinge loss on its own, the combination of the two yields the best performance on average, the difference being more pronounced in the more complex ArtCap domain.

## 12. Impact of False Negatives

In an open-domain setting, a given query $x$ may have multiple valid image matches $Y$, even though our evaluation setup only designates a single ground-truth. For instance, queries from COCO may have equally valid counterparts in closely related datasets such as Flickr, which are then characterized as *false negatives*. This ambiguity makes evaluation nuanced. Nevertheless, EFSA consistently improves Recall@1 with respect to the designated ground-truth targets, suggesting that the false negative issue, while possible, is not ubiquitous. In cases where it does interfere with the interpretation of Recall@1, we can additionally rely on Recall@5 and Recall@10, which reflect whether the ground-truth is still being successfully promoted to the top ranks despite the presence of potential false negatives.




(a) Qualitative examples from ArtCap.




(b) Qualitative examples from TextCap.

Figure 7. Qualitative comparison between EFSA and zero-shot CLIP on the ArtCap (top teo examples) and TextCap (bottom two examples) datasets in the single-domain setting. Green-framed images indicate the ground-truth for each text query, displayed on top. EFSA effectively re-ranks the ground-truth images to the top rank, outperforming zero-shot CLIP. On the right, the synthetic caption for each image is provided, as used for episodic few-shot adaptation.