CaPCL: Caption-Preserved Continual Learning for Text-to-Image Retrieval

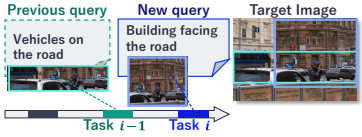
Naoya Sogi[1], Ren Ohkubo[2,3], Takashi Shibata[1], Makoto Terao[1], Yusuke Hosoya[4], and Takayuki Okatani[4,5]

[1] Cyber Physical Intelligence Research Labs., NEC Corporation, Kanagawa, Japan
[2] University of Tsukuba, Ibaraki, Japan
[3] National Institute of Advanced Industrial Science and Technology, Ibaraki, Japan
[4] Graduate School of Information Sciences, Tohoku University, Miyagi, Japan
[5] RIKEN Center for AIP, Tokyo, Japan
naoya-sogi@nec.com, ookubo.0720@aist.go.jp, t.shibata@ieee.org, m-terao@nec.com, {yhosoya, okatani}@vision.is.tohoku.ac.jp

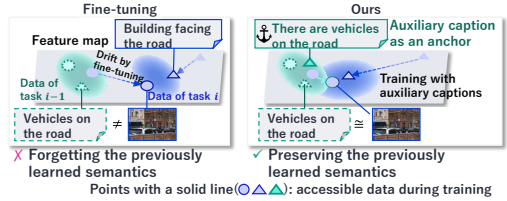
Abstract. Continual learning for text-to-image retrieval is essential for adapting retrieval systems to evolving user interests and emerging concepts. While existing methods mitigate forgetting through image-text similarity distillation, they overlook a critical real-world scenario: new textual queries often describe different semantic facets of images from previously seen visual domains, a phenomenon we term "unbalanced cross-modal shift". This shift arises from the inherent polysemy of visual content and is strongly associated with severe forgetting. To study this phenomenon, we introduce two diagnostic metrics: image domain overlap measured via principal subspace similarity, and textual query divergence quantified through image-conditioned query probability. Our empirical analysis shows that unbalanced cross-modal shift significantly amplifies forgetting, with existing methods struggling to retain image semantics under such conditions. To address this challenge, we propose CaPCL (Caption-Preserved Continual Learning), which regularizes models to maintain consistent caption generation for previously learned content, thereby preserving multi-faceted image semantics. By leveraging auxiliary captions as semantic anchors, CaPCL delivers more robust training signals than contrastive similarity alone. We construct a benchmark integrating five heterogeneous datasets to enable comprehensive evaluation. Extensive experiments demonstrate that CaPCL consistently outperforms state-of-the-art methods in both knowledge retention and adaptation.

1 Introduction

Text-to-Image (T2I) retrieval is a fundamental capability for applications ranging from multimedia search engines to e-commerce platforms [3, 9, 10, 19, 30]. Vision-Language Models (VLMs) represented by CLIP [40], which learn a joint embedding space through large-scale contrastive pretraining, have significantly



(a) Illustration of image polysemy in continual T2I retrieval. An image initially associated with "vehicles on the road" can later be queried as "building facing the road," highlighting how unbalanced cross-modal shift occurs when textual semantics evolve while visual domains remain stable.



(b) Conceptual diagram of CaPCL. By generating auxiliary captions that capture previously learned semantics and regularizing the model to preserve these captions, CaPCL retains multi-faceted semantics.

Fig. 1: (a) Motivating example of unbalanced cross-modal shift due to image polysemy and (b) Overview of CaPCL.

improved retrieval by accurately aligning both modalities. More recently, multi-task VLMs such as BLIP-2 [24] and SigLIP2 [46] have further enhanced these representations by incorporating caption generation, leading to superior retrieval performance.

Real-world retrieval systems must continually adapt to new concepts and evolving user preferences while preserving previously acquired knowledge. To address this, Wang *et al.* [47] formalized a continual learning setting where a foundational VLM (original model), already proficient in general T2I tasks, is incrementally trained on new image-text pairs while maintaining its existing capabilities. Specialized methods like ModX [37] and DKR [7] address this through similarity matrix distillation, which regularizes the model to preserve image-text similarity predictions computed before training on each new task, thereby stabilizing the embedding space and mitigating forgetting.

However, these existing studies overlook a critical real-world scenario. This is the occurrence of a distorted distribution shift unique to T2I retrieval—which we term "unbalanced cross-modal shift"—where the visual domain of images remains stable while the text distribution of queries changes dramatically. This phenomenon arises from "image polysemy," where a single image inherently contains multiple semantic facets. As illustrated in Fig. 1a, it occurs through the transition of user interests or intent, such as when an image initially retrieved with the query "vehicles on the road" is later sought from a different perspective as "building facing the road". In the real world, user needs are volatile and evolve more rapidly than visual information; thus, such unbalanced shifts are common.

In this paper, we first analyze the impact of unbalanced cross-modal shift on T2I retrieval. Specifically, we introduce two metrics to capture distributional changes across task transitions for each modality: (1) image domain overlap based on principal subspace similarity, and (2) textual query divergence. Through experiments on diverse datasets, we demonstrate that unbalanced cross-modal shift is associated with greater forgetting when a model is adapted to new T2I tasks using existing methods. This occurs because conventional approaches

implicitly assume that both image and text modalities shift simultaneously, failing to preserve the inherent multi-faceted semantics of images.

Next, to address this challenge, we propose CaPCL (**Ca**ption-**P**reserved **C**ontinual **L**earning), a method that leverages the captioning capabilities of modern retrieval-oriented VLMs to preserve multi-faceted image semantics, as outlined in Fig. 1b. Given an original VLM—a pre-trained retrieval-oriented model with captioning ability—CaPCL generates an auxiliary caption for each new image using the pre-update model. This caption reflects the model's current semantic interpretation of the image and may capture aspects beyond those expressed by the new query. While updating the model on new image-text pairs, CaPCL regularizes it to reproduce these captions, thereby preserving previously learned semantics. Importantly, caption preservation constrains the semantic interpretation of an image itself, whereas conventional similarity-preservation approaches constrain only its relationship with specific text queries. Consequently, under unbalanced cross-modal shift, similarity-based methods may allow the image representation to drift toward newly introduced textual semantics, whereas CaPCL anchors the model's prior interpretation of each image and mitigates catastrophic forgetting.

We further construct a comprehensive benchmark that integrates five heterogeneous datasets, reflecting realistic scenarios. Extensive experiments demonstrate that CaPCL substantially outperforms state-of-the-art methods in both knowledge retention and adaptation, validating the effectiveness of CaPCL [1].

In summary, our main contributions are:

- We identify and empirically characterize a critical yet overlooked scenario in continual T2I retrieval: unbalanced cross-modal shift, where high image domain overlap and large textual query divergence are associated with severe forgetting.
- We propose CaPCL, a novel continual learning method that leverages caption-based regularization to preserve multi-faceted semantic knowledge, along with optional caption-aware extensions.
- We construct a continual retrieval benchmark that integrates five heterogeneous datasets to enable comprehensive evaluation.
- We provide extensive experiments demonstrating that CaPCL outperforms state-of-the-art baselines in both retention and adaptation metrics.

2 Related Work

- *Text-to-Image Retrieval:* Text-to-Image (T2I) retrieval aims to retrieve relevant images based on textual queries, a fundamental task in vision-language research [3, 9, 19]. Early methods employed region-based features to model fine-grained correspondences between image regions and textual phrases [5, 23, 27, 35, 39, 45]. This research field was transformed by large-scale pre-trained models

[1] Code will be released at `https://github.com/NEC-N-SOGI/CaPCL`.

that learn aligned image and text embeddings through contrastive learning on massive datasets [16, 20, 25, 26, 28, 40, 44, 53, 54]. These models establish shared embedding spaces that underpin modern retrieval systems.

Multi-task learning further enhances VLM performance by combining contrastive training with generative objectives. Models like BLIP-2 [24] and SigLIP2 [46] integrate image captioning, enriching representations with semantic knowledge. Despite these advances in retrieval performance, continual adaptation remains a significant challenge.

- *Continual Learning for T2I Retrieval:* Continual learning aims to enable models to learn sequentially from new data while retaining performance on previously learned tasks. In image classification, methods such as Elastic Weight Consolidation (EWC) [22] and Learning without Forgetting (LwF) [29] employ regularization to prevent catastrophic forgetting. More recently, parameter-efficient approaches, such as visual prompt tuning [17] and LoRA [15], have been developed for adapting large foundation models, including vision-language models.

In the context of T2I retrieval, continual learning is critical as user interests and data distributions evolve over time. The seminal work [47] established the continual learning setting for T2I retrieval and demonstrated the effectiveness of EWC-based regularization. Subsequent methods, such as ModX [37] and DKR [7], develop retrieval-specific strategies by distilling image-text similarity matrices. These approaches compute similarity matrices using the pre-update model before training, then minimize the deviation of the updated model's predictions from these reference similarities. Recent work C2MR [55] further utilizes feature distillation to preserve the embedding space, and C-CLIP [32] combines LoRA-based adaptation with embedding-space knowledge consolidation to retain prior representations. These strategies stabilize the embedding space and help maintain cross-modal associations.

However, these methods focus exclusively on new image-query pairs, neglecting the multi-faceted semantics that images may possess beyond what new queries capture. Consequently, when faced with an unbalanced cross-modal shift, these methods struggle to retain prior semantic associations. To address this limitation, our method leverages the generative capabilities of retrieval-oriented VLMs to preserve image semantics through caption-based regularization.

3 Analysis of Forgetting in Text-to-Image Retrieval

This section analyzes forgetting behavior in continual T2I retrieval, with a focus on scenarios involving unbalanced cross-modal shifts. We first describe our settings, then introduce two metrics to quantify unbalanced cross-modal shift, and finally present empirical results demonstrating its impact on forgetting.

3.1 Settings

We evaluate how continual learning affects retrieval accuracy on Flickr30K [52], a widely used T2I benchmark. As the original model, we use the BLIP-2 [24] model,

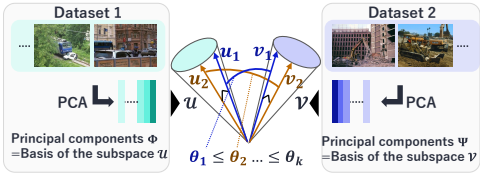


Fig. 2: Illustration of principal angles between two subspaces $\mathcal{U}$ and $\mathcal{V}$. The first principal angle θ_1 represents the smallest angle between any pair of unit vectors from the two subspaces, while subsequent angles capture independent directions. Smaller angles indicate greater overlap between the distributions.

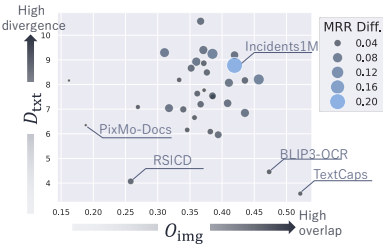


Fig. 3: Relationship among O_{img}, D_{txt}, and forgetting behavior measured by MRR on Flickr30K. Larger points indicate greater forgetting.

which is trained on COCO and exhibits strong performance on Flickr30K. We train this model on each dataset and measure the performance change relative to the original model using MRR (Mean Reciprocal Rank) [6].

Our training data comprises five large-scale datasets spanning diverse domains: Incidents1M [49], BLIP3-OCR [50], RSICD [36], TextCaps [43], and PixMo-Docs [8] (detailed in Sec. 5). To broaden evaluation coverage, we augment these with 27 datasets sampled from ImageNet-1K [41], yielding a total of 32 datasets. The ImageNet-based datasets vary in granularity: 5 datasets contain 2 classes, 12 contain 5 classes, and 10 contain 10 classes, each with 1,000 images per class. For these datasets, we use class names as textual queries.

After training the original BLIP-2 model on each dataset individually, we evaluate MRR on the Flickr30K test set. Additional settings and implementation details are provided in the supplementary material (Sec. A.1).

3.2 Quantifying Unbalanced Cross-Modal Shift

To quantify the degree of unbalanced cross-modal shift, we introduce two complementary metrics: image domain overlap and textual query divergence.

- *Image Domain Overlap:* We measure the overlap between image domains using the similarity of their principal component subspaces [12, 51]. The use of principal subspaces and their principal angles enables us to capture even partial overlaps between distributions, even when their means are separated. For instance, maximum mean discrepancy [11] cannot capture partial overlaps when the means are distant, unlike our approach.

We extract image features from both the Flickr30K test set and the training datasets using the original BLIP-2 image encoder. For each dataset, we perform principal component analysis (PCA) to obtain the top k principal components (PCs), which define a k-dimensional subspace characterizing the image distribution. We automatically set k using the cumulative explained variance ratio with a threshold of 0.95 for each dataset. This threshold enables the capture of the structure of the data distributions.

To quantify the similarity between these subspaces, we employ principal angles $\{\theta_i\}_{i=1}^{k}$, a standard measure of subspace alignment [1, 14]. As shown in Fig. 2, the i-th principal angle θ_i is recursively defined as:

$$\cos\theta_i = \max_{\mathbf{u}_i \in \mathcal{U}} \max_{\mathbf{v}_i \in \mathcal{V}} \mathbf{u}_i^T \mathbf{v}_i, \text{s.t.} \quad \|\mathbf{u}_i\| = \|\mathbf{v}_i\| = 1, \mathbf{u}_i^T \mathbf{u}_j = \mathbf{v}_i^T \mathbf{v}_j = 0, \forall j \neq i, \quad (1)$$

where $\mathcal{U}$ and $\mathcal{V}$ denote the two subspaces. Each value $\cos\theta_i$ can be computed by utilizing singular value decomposition [1, 14].

We define the image domain overlap as the proportion of strongly aligned principal directions:

$$O_{\text{img}} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(\cos\theta_i > 0.95), \quad (2)$$

where $\mathbb{1}(\cdot)$ is the indicator function, equal to 1 if the condition is true and 0 otherwise. The condition $\cos\theta_i > 0.95$ indicates that $\mathbf{u}_i, \mathbf{v}_i$ are closely aligned, suggesting significant overlap. Thus, O_{img} values close to 1 indicate high overlap. The detailed procedure is provided in the supplementary material (Sec. A.2).

- *Textual Query Divergence:* We measure textual query divergence using the captioning capability of a VLM. For each image-text pair (I, T) in the training datasets, we compute the log-probability of the text T conditioned on I:

$$\ell(T, I) = \frac{1}{|T| - 1} \sum_{t=2}^{|T|} \log p_{\text{VLM}}(T_t \mid I, T_{<t}), \quad (3)$$

where T_t denotes the t-th token of the text, and p_{VLM} is a token-level probability computed by a VLM. The textual query divergence is defined as the average negative log-probability across all pairs in a training dataset:

$$D_{\text{txt}} = \frac{-1}{N} \sum_{i=1}^{N} \ell(T_i, I_i), \quad (4)$$

where N is the number of image-text pairs. Higher D_{txt} indicates a greater divergence from previously learned semantics. We use the same BLIP-2 model for computing this metric.

3.3 Results

Figure 3 illustrates the relationship between our two metrics and the observed forgetting, as measured by the drop in MRR on Flickr30K after training. The results show that severe forgetting tends to occur when image domains overlap significantly and textual queries diverge substantially.

For example, Incidents1M exhibits the most severe forgetting (-0.20), with high image domain overlap (0.42) and substantial textual query divergence (8.9). This suggests that knowledge about previously learned image domains can be

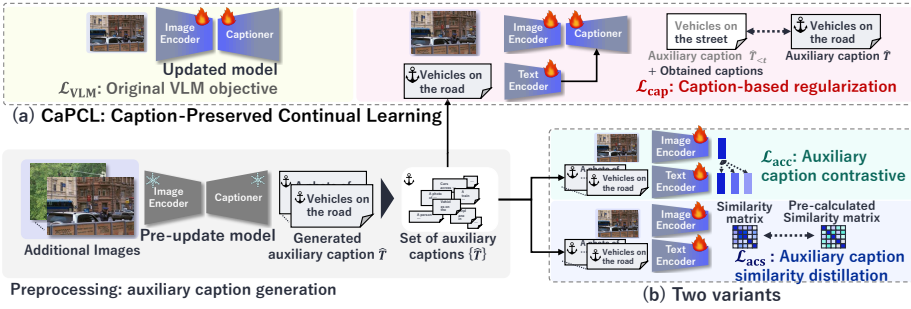


Fig. 4: (a) Overview of CaPCL. For each incoming pair (I, T), an auxiliary caption $\hat{T}$ is generated using the pre-update model's captioning head, which captures the model's prior understanding of the image. The model is then trained with caption-based regularization to maintain consistent caption generation. (b) Overview of the two optional extensions that further leverage auxiliary captions: Auxiliary Caption Contrastive (ACC) loss and Auxiliary Caption Similarity (ACS) distillation loss.

overwritten when new textual queries emphasize substantially different semantics, as will be exemplified in Fig. 6. In contrast, when image domain overlap is low (<0.30), even significant textual query divergence results in relatively minor forgetting. Similarly, high image domain overlap combined with low textual query divergence (<6.0) also leads to limited forgetting. The joint score $O_{\text{img}} \times D_{\text{txt}}$ correlates more strongly with forgetting (ρ=0.59) than either O_{img} or D_{txt} alone (ρ=0.33 and 0.45, respectively), and this correlation is robust to the hyperparameters of O_{img} (the PCA explained-variance and cosine thresholds), as reported in Sec. C.2 of the supplementary material. These findings confirm that unbalanced cross-modal shift is strongly associated with severe forgetting in continual T2I retrieval.

To address this challenge, we propose CaPCL, which uses caption-based regularization to preserve previously learned semantics.

4 Proposed Method

This section details CaPCL. Conventional methods stabilize the embedding space using similarity information from image-text pairs, which constrains only pairwise image–text relationships and may fail to retain other semantic aspects of an image. Under unbalanced cross-modal shift, this can cause the image representation to drift toward newly introduced textual semantics. To address this limitation, CaPCL generates an auxiliary caption for each image using the pre-update model and uses it as a regularization signal during training. These captions capture semantic aspects of the image beyond those expressed by the new query. In this way, CaPCL constrains the semantic interpretation of each image itself, rather than only its similarity to specific text queries.

We first describe the core caption-based regularization mechanism, then present optional extensions, and discuss how to extend our method to retrieval-oriented VLMs without captioning capabilities.

In the following, the *pre-update model* denotes the frozen model snapshot just before training on a new dataset begins; when training on the first dataset, this is the original model. The *updated model* denotes the model being trained on that dataset.

4.1 Caption-Based Regularization

- *Overview:* Figure 4 illustrates the overall framework of CaPCL. Given an original VLM pre-trained for T2I retrieval with captioning capabilities, CaPCL continually adapts to new image-text pairs while preserving prior knowledge. For each incoming pair (I, T), we generate an auxiliary caption $\hat{T}$ using the captioning capability of the pre-update model, capturing the model's semantic interpretation of the image before adaptation. CaPCL then employs caption-based regularization to encourage the updated model to maintain consistent caption generation, thereby mitigating catastrophic forgetting.

- *Detail:* CaPCL combines the original VLM loss $\mathcal{L}_{\text{VLM}}$ with caption-based regularization $\mathcal{L}_{\text{cap}}$:

$$\mathcal{L} = \mathcal{L}_{\text{VLM}} + \mathcal{L}_{\text{cap}}, \tag{5}$$

where $\mathcal{L}_{\text{VLM}}$ includes the contrastive loss and other objectives used by the original model (e.g., caption generation loss for new queries). Here, $\mathcal{L}_{\text{VLM}}$ enables adaptation to new data, while $\mathcal{L}_{\text{cap}}$ allows the model to retain previously learned semantics.

For each new image-text pair (I, T), we generate an auxiliary caption $\hat{T}$ using the pre-update model through nucleus sampling [13]. This auxiliary caption captures aspects beyond what the new query represents. The caption-based regularization loss is defined as the negative log-likelihood of this auxiliary caption $\hat{T}$:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{|\hat{T}|} \log p_{\text{VLM}}(\hat{T}_t \mid I, \hat{T}_{<t}), \tag{6}$$

where $\hat{T}_t$ denotes the t-th token of the auxiliary caption, and $\hat{T}_{<t}$ represents all preceding tokens. This loss encourages the updated model to generate captions consistent with its prior understanding. By integrating this caption-based regularization $\mathcal{L}_{\text{cap}}$ with the original learning objectives $\mathcal{L}_{\text{VLM}}$ as in Eq. (5), CaPCL effectively mitigates catastrophic forgetting while enabling adaptation to new data.

4.2 Variations in Auxiliary Caption Utilization

The auxiliary captions introduced above can be utilized in multiple ways during training. In addition to the caption-based regularization described above, they can also be used in additional training losses, such as those used in prior T2I retrieval-specific continual learning approaches [7, 37, 55]. These losses can be viewed as optional extensions that complement the base formulation. As described later, they also enable the use of CaPCL with retrieval-oriented VLMs that do not possess captioning capabilities.

- *Auxiliary Caption Contrastive (ACC) Loss:* We can incorporate a contrastive objective between images and their auxiliary captions:

$$\mathcal{L}_{\text{acc}} = -\log \frac{\exp(\text{sim}(I, \hat{T})/\tau)}{\sum_{T' \in \mathcal{T}} \exp(\text{sim}(I, T')/\tau)} - \log \frac{\exp(\text{sim}(\hat{T}, I)/\tau)}{\sum_{I' \in \mathcal{I}} \exp(\text{sim}(\hat{T}, I')/\tau)}, \quad (7)$$

where sim denotes cosine similarity, τ is a temperature, $\mathcal{T}$ and $\mathcal{I}$ are the sets of auxiliary captions and images in the batch, respectively. This loss encourages the model to maintain strong associations between images and their auxiliary captions, reinforcing the preservation of multi-faceted semantics.

- *Auxiliary Caption Similarity (ACS) Distillation Loss:* Another extension involves distilling the similarity structure between images and auxiliary captions from the pre-update model, similar to prior similarity matrix distillation methods [7, 37, 55]. For each image I, we compute similarity scores with all auxiliary captions in the batch using both the pre-update and the updated models:

$$s_{\text{pre}}(I, \hat{T}_i) = \text{sim}_{\text{pre}}(I, \hat{T}_i), \;\; s_{\text{cl}}(I, \hat{T}_i) = \text{sim}_{\text{cl}}(I, \hat{T}_i), \quad (8)$$

where $\hat{T}_i$ denotes the i-th auxiliary caption in the batch, and sim_{pre} and sim_{cl} represent similarity computations using the pre-update and the updated models, respectively.

We apply temperature-scaled softmax to these similarity scores and compute Kullback-Leibler (KL) divergence between the resulting distributions:

$$\mathcal{L}_{\text{acs}}^{img} = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \text{KL}\Big(\sigma(\frac{s_{\text{cl}}(I, \hat{T}_i)}{\tau}) \| \sigma(\frac{s_{\text{pre}}(I, \hat{T}_i)}{\tau})\Big), \quad (9)$$

where $\sigma(\cdot)$ denotes the softmax function. Similarly, we compute the KL loss from the caption perspective:

$$\mathcal{L}_{\text{acs}}^{cap} = \frac{1}{|\mathcal{T}|} \sum_{\hat{T} \in \mathcal{T}} \text{KL}\Big(\sigma(\frac{s_{\text{cl}}(\hat{T}, I_i)}{\tau}) \| \sigma(\frac{s_{\text{pre}}(\hat{T}, I_i)}{\tau})\Big). \quad (10)$$

The overall ACS distillation loss is then given by:

$$\mathcal{L}_{\text{acs}} = \mathcal{L}_{\text{acs}}^{img} + \mathcal{L}_{\text{acs}}^{cap}. \quad (11)$$

This loss directly preserves the similarity structure between images and auxiliary captions, further mitigating forgetting.

- *Overall Loss:* When using ACC loss $\mathcal{L}_{\text{acc}}$ and ACS distillation loss $\mathcal{L}_{\text{acs}}$ in addition to the caption-based regularization $\mathcal{L}_{\text{cap}}$, the overall loss function becomes:

$$\mathcal{L} = \mathcal{L}_{\text{VLM}} + \mathcal{L}_{\text{cap}} + \mathcal{L}_{\text{acc}} + \mathcal{L}_{\text{acs}}. \tag{12}$$

As demonstrated in Sec. 5, our method, which only uses $\mathcal{L}_{\text{cap}}$ and $\mathcal{L}_{\text{VLM}}$, outperforms conventional methods, while integrating these variations further enhances performance.

- *Extension to Retrieval-Only VLMs:* CaPCL relies on auxiliary captions generated by the pre-update VLM, which we refer to as *self-generated* captions. However, some retrieval-oriented VLMs lack inherent captioning capabilities. To extend our approach to such models, we utilize an external captioning model, such as Qwen2.5-VL [2], to generate auxiliary captions. These auxiliary captions capture common aspects of the image content, potentially differing from the new textual query T and thereby provide multi-faceted semantic information. We then apply the optional extensions described above, which do not require the original VLM's captioning head. This approach enables us to apply our method to retrieval-oriented VLMs without captioning capabilities.

5 Experiments

This section validates the effectiveness of CaPCL. We first describe the setup and then present the main results, followed by ablation studies that analyze the contributions of different components. Additional details are provided in the supplementary material (Sec. B). Further ablation studies on the number of auxiliary captions, caption quality and diversity, and the weight of caption-based regularization are provided in the supplementary material (Sec. C).

5.1 Experimental Setup

- *Datasets and Evaluation Protocol:* We evaluate CaPCL on a comprehensive benchmark constructed from five datasets: Incidents1M [48,49], BLIP3-OCR [50], RSICD [36], TextCaps [43], and PixMo-Docs [8]. Each dataset presents unique challenges and varying degrees of distribution shifts, enabling a thorough assessment. For instance, Incidents1M contains images related to real-world incident names, which may differ significantly from prior queries.

A retrieval model is continually trained on each dataset in sequence. To evaluate robustness across different task sequences, we consider four training orders denoted by dataset initials: 1) *IBRTP* (Incidents1M → BLIP3-OCR → RSICD→ TextCaps → PixMo), 2) *PTRBI*, 3) *BTIRP*, and 4) *PRITB*. After training on each dataset, we evaluate retrieval performance on all previously seen datasets and the Flickr30K test set [18,52] to measure both adaptation and retention. As many VLMs exhibit strong zero-shot performance on Flickr30K, this evaluation also assesses how continual learning affects the zero-shot performance.

Table 1: Comparative results. The top four blocks correspond to different data addition orders. The left column indicates the order of dataset addition. The bottom block shows the average results across all orders. The best and second-best results are highlighted in **bold** and underlined, respectively. The original BLIP-2 model without additional learning achieves AA of 0.5409.

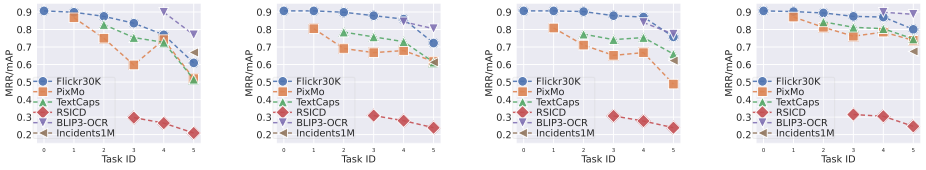
		FT	EWC [22]	ModX [37]	DKR [7]	C2MR [55]	CaPCL
IBRTP	AA ↑	0.6830	0.6874	0.6518	0.6656	0.6525	**0.7019**
	FM ↓	0.0791	0.0742	0.0764	0.0748	0.0696	**0.0554**
	HM ↑	0.7843	0.7890	0.7642	0.7742	0.7671	**0.8054**
PTRBI	AA ↑	0.5277	0.5219	0.6016	0.5327	0.5902	**0.6690**
	FM ↓	0.2674	0.2730	0.1303	0.1814	0.1432	**0.0965**
	HM ↑	0.6135	0.6076	0.7112	0.6454	0.6990	**0.7688**
BTIRP	AA ↑	0.6951	0.6921	0.6641	0.6731	0.6339	**0.7058**
	FM ↓	0.0647	0.0690	0.0635	0.0587	0.0953	**0.0502**
	HM ↑	0.7975	0.7940	0.7771	0.7850	0.7455	**0.8098**
PRITB	AA ↑	0.6780	0.6827	0.6374	0.6483	0.6433	**0.7000**
	FM ↓	0.0851	0.0802	0.0860	0.0851	0.0741	**0.0594**
	HM ↑	0.7789	0.7837	0.7511	0.7589	0.7591	**0.8027**
Average	AA ↑	0.6459	0.6460	0.6387	0.6300	0.6300	**0.6942**
	FM ↓	0.1241	0.1241	0.0890	0.1000	0.0955	**0.0654**
	HM ↑	0.7435	0.7436	0.7509	0.7409	0.7427	**0.7966**

- *Evaluation Metrics:* For evaluating retrieval performance, we use Mean Reciprocal Rank (MRR) [6] and Mean Average Precision (mAP) [42]. MRR is used for Flickr30K, BLIP3-OCR, RSICD, TextCaps, and PixMo, while mAP is used for Incidents1M, as the first five datasets involve one-to-one correspondence between texts and images, whereas Incidents1M involves one-to-many correspondence.

For continual learning evaluation, we report Average Accuracy (AA) [4, 33] as a measure of adaptation to new data and Forgetting Measure (FM) [4] as a measure of degradation of existing knowledge. We use the harmonic mean (HM) of AA and $1 - \text{FM}$ as an overall evaluation metric.

- *Implementation Details:* We primarily use a BLIP-2 model fine-tuned for retrieval on COCO [31] as the original model, as it is the state-of-the-art VLM for T2I retrieval. For $\mathcal{L}_{\text{VLM}}$, we utilize the contrastive loss between image and text embeddings, along with the language modeling loss for generating new textual queries, as in the original BLIP-2 training. We use the Adam optimizer [21] with a learning rate of 1e-4, employing a 10% warmup and the cosine annealing schedule [34]. The trainable parameters include all parameters of the BLIP-2's Q-Former, which connects images and text, as well as the captioning head.

We implement all methods using 8 NVIDIA L40S GPUs with PyTorch [38], and train for 10 epochs on each dataset.



(a) FT **(b)** ModX **(c)** C2MR **(d)** CaPCL

Fig. 5: Performance transitions for the *PTRBI* order. The horizontal axis shows the number of datasets learned, and zero means the original BLIP-2 model. The metrics for each dataset are reported after learning on its training split.

CaPCL has two hyperparameters: the temperature τ used in contrastive learning and similarity distillation, and the number of auxiliary captions generated per image. We set $\tau = 0.0295$, the same as that in BLIP-2 [24], and generate five auxiliary captions per image using nucleus sampling with $p = 0.9$. The loss $\mathcal{L}_{\text{cap}}$ is computed by averaging the negative log-likelihoods of all captions. Auxiliary caption generation is a one-time preprocessing step per dataset, and its cost is small relative to training.[2]

We also conduct experiments using the SigLIP2 [46] giant (384) model as the original model to validate the generality of our method. For $\mathcal{L}_{\text{VLM}}$, we use the sigmoid-based contrastive loss as in the original SigLIP2 training. The other training settings are identical to those for BLIP-2.

- *Baseline Methods:* We compare CaPCL against four state-of-the-art continual learning methods for T2I retrieval: ModX [37], DKR [7], and C2MR [55], as well as a standard baseline: EWC [22]. For a fair comparison, we implement all baselines using the same BLIP-2 model and training settings as CaPCL, including the same VLM-native losses and trainable modules. We additionally compare against C-CLIP [32], a strong LoRA-based retention method, with a controlled comparison provided in the supplementary material (Sec. C.1). We also include a fine-tuning (FT) baseline that learns using only the standard VLM loss $\mathcal{L}_{\text{VLM}}$. Please refer to the supplementary material (Sec. B.4) for details of the implementation of each baseline method.

5.2 Results and Discussion

Overall Performance Comparison Table 1 and Fig. 5 present the performance comparison. The proposed CaPCL consistently outperforms all baselines across various training orders, achieving the best metrics in terms of AA, FM, and HM. These results validate that our method effectively balances adaptation to new data while retaining previously learned knowledge.

[2] For 50K image-text pairs with five captions per image, generation takes approximately 35 minutes on 8 L40S GPUs, compared with approximately 160 minutes for 10-epoch training.

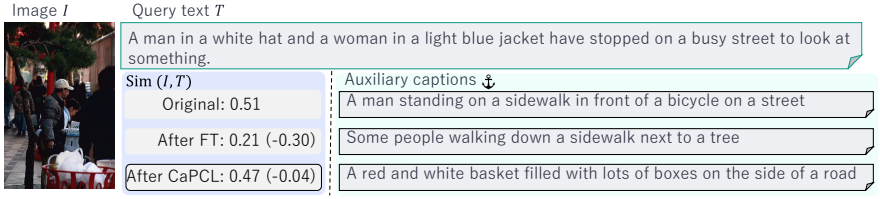


Fig. 6: Example of similarity changes and auxiliary captions. The similarity changes before (original) and after (FT/CaPCL) training on Incidents1M are shown.

Table 2: Ablation study on variants of auxiliary caption utilization. $\mathcal{L}_{cap}$ denotes caption-based regularization (CaPCL), $\mathcal{L}_{acc}$ denotes auxiliary caption contrastive loss, and $\mathcal{L}_{acs}$ denotes auxiliary caption similarity distillation loss. The results are averaged over all training orders. The best performance by the baseline method is achieved by ModX, with AA of 0.6387, FM of 0.0890, and HM of 0.7509.

$\mathcal{L}_{cap}$	✓			✓	✓	✓
$\mathcal{L}_{acc}$		✓		✓		✓
$\mathcal{L}_{acs}$			✓		✓	✓
AA ↑	**0.6942**	0.6666	0.6301	0.6894	0.6658	<u>0.6941</u>
FM ↓	<u>0.0654</u>	0.0929	0.1401	<u>0.0654</u>	0.0977	**0.0596**
HM ↑	<u>0.7966</u>	0.7684	0.7272	0.7935	0.7662	**0.7987**

Among the baseline methods, ModX shows relatively good forgetting performance (FM of 0.0890) but suffers from lower adaptation (AA of 0.6387). This suggests that similarity matrix distillation overly constrains the model's ability to learn effective features for new data. In contrast, CaPCL preserves semantic knowledge without directly constraining feature representations, allowing for more effective adaptation to new data. This leads to a better balance between retention and adaptation.

The performance varies across different training orders. For instance, the *PTRBI* order exhibits the highest forgetting across all methods, suggesting that this particular sequence presents more challenging distribution shifts. However, CaPCL maintains robust performance even under this challenging order, achieving an FM of 0.0965, which is substantially lower than all baselines. This consistency across the training orders demonstrates the robustness of our caption-based regularization approach.

Figure 6 illustrates an example of similarity changes before and after training on Incidents1M and auxiliary captions. The figure shows that fine-tuning (FT) leads to a significant drop in similarity for the relevant image-text pair, indicating forgetting. In contrast, CaPCL maintains higher similarity for the pair, demonstrating its effectiveness in preserving prior knowledge. The auxiliary captions capture essential semantic details, which guide the updated model to retain relevant information during continual learning.

Table 3: Effectiveness of self-caption generation. Comparison of using self-generated captions from the pre-update BLIP-2 model, captions generated by Qwen2.5-VL, and a combination of both as auxiliary captions for caption-based regularization.

	Self	Qwen	Both
AA ↑	<u>0.6942</u>	0.6872	**0.6972**
Average FM ↓	<u>0.0654</u>	0.0754	**0.0613**
HM ↑	<u>0.7966</u>	0.7884	**0.8001**

Table 4: Effectiveness of auxiliary captions on the SigLIP2 model.

Auxiliary captions are generated by Qwen2.5-VL and used to compute the auxiliary caption contrastive loss $\mathcal{L}_{acc}$. The results are averaged over all training orders.

	FT	ModX [37]	Ours ($\mathcal{L}_{acc}$)
AA ↑	0.5262	0.4786	**0.5939**
FM ↓	0.1905	0.1580	**0.1086**
HM ↑	0.6377	0.6082	**0.7129**

Effectiveness of Caption-Based Regularization This subsection presents an ablation study on the variants of auxiliary caption utilization, as introduced in Sec. 4.2.

Table 2 shows the results of different combinations of the proposed loss components. The comparison of using each component individually shows that CaPCL is the most effective. Comparing individual components, $\mathcal{L}_{\text{cap}}$ alone achieves the strongest performance, outperforming ModX in both adaptation and retention. This demonstrates that caption preservation effectively maintains semantic knowledge while allowing flexible adaptation to new data, as discussed in the previous subsection. The ACC loss $\mathcal{L}_{\text{acc}}$ alone shows better adaptation and comparable forgetting prevention compared to ModX, indicating that contrastive learning with auxiliary captions helps retain prior knowledge. The combination of all three components achieves the best performance, demonstrating that these auxiliary caption utilizations are complementary.

Effectiveness of Self-Caption Generation CaPCL $\mathcal{L}_{\text{cap}}$ relies on self-generated captions to capture the model's semantic understanding. This subsection evaluates the effectiveness of using self-generated captions. To this end, we generate auxiliary captions using the Qwen2.5-VL 7B model [2], a state-of-the-art vision-language model, and train the BLIP-2 model with these captions as an alternative auxiliary caption source. Additionally, we evaluate the combination of both self-generated and Qwen2.5-VL-generated auxiliary captions, i.e., we calculate the caption-based loss for both sets of captions and sum them up.

Table 3 shows the results of this comparison. The results show that self-generated captions outperform those generated by Qwen2.5-VL. This highlights the importance of capturing and preserving the knowledge embedded in the pre-update model. However, even when using Qwen-generated captions, our method still outperforms existing baselines. This suggests that high-quality VLMs, such as Qwen2.5-VL, possess rich general knowledge about images, and leveraging this knowledge through caption preservation is beneficial for mitigating forgetting. Furthermore, our findings suggest that CaPCL has the potential to easily leverage future VLMs for retrieval, even those without captioning capabilities. The

combination of both caption sources yields the best performance, indicating that they provide complementary semantic information for knowledge preservation.

Generality Across Vision-Language Architectures To examine whether auxiliary captions are useful beyond BLIP-2, we apply our framework to SigLIP2 [46]. Since SigLIP2's captioning head is not publicly available, this experiment does not evaluate $\mathcal{L}_{\text{cap}}$ itself; instead, it uses Qwen2.5-VL-generated captions with the ACC loss.

Table 4 shows that even with this different architecture, our method consistently outperforms baselines. While SigLIP2's overall retrieval performance is slightly lower than BLIP-2 due to architectural and training differences, the relative improvements indicate that auxiliary captions are beneficial across different vision-language architectures.

6 Conclusion

In this paper, we identified unbalanced cross-modal shift as a critical challenge in continual text-to-image retrieval, where high image domain overlap combined with significant textual query divergence leads to severe forgetting. To address this, we proposed CaPCL, a novel continual learning method that leverages caption-based regularization to preserve multi-faceted semantic knowledge. Extensive experiments on a comprehensive benchmark constructed from five diverse datasets demonstrated that CaPCL outperforms state-of-the-art methods in both retention and adaptation metrics. Our method relies on the quality of the captioning head, and what knowledge is preserved is implicitly determined by the captioning head rather than explicitly controlled. Future work could extend our approach to prioritize the preservation of explicitly specified knowledge when such information is available. We believe that our work will stimulate further research into the nature of incoming data in continual learning on T2I retrieval and inspire the development of methods tailored to the evolving capabilities of modern vision-language models.

References

1. Afriat, S.N.: Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. In: Mathematical proceedings of the Cambridge philosophical society. vol. 53, pp. 800–816 (1957)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-Text Retrieval: A Survey on Recent Research and Development. In: IJCAI. pp. 5410–5417 (2022)
4. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: ECCV. pp. 532–547 (2018)

5. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120 (2020)
6. Craswell, N.: Mean Reciprocal Rank, pp. 1703–1703 (2009)
7. Cui, Z., Peng, Y., Wang, X., Zhu, M., Zhou, J.: Continual vision-language retrieval via dynamic knowledge rectification. In: AAAI. vol. 38, pp. 11704–11712 (2024)
8. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR. pp. 91–104 (2025)
9. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS. vol. 26 (2013)
10. Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., Natarajan, P.: Fashionvlp: Vision language transformer for fashion retrieval with feedback. In: CVPR. pp. 14105–14115 (2022)
11. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. JMLR **13**(1), 723–773 (2012)
12. Hamm, J., Lee, D.D.: Grassmann discriminant analysis: a unifying view on subspace-based learning. In: ICML. pp. 376–383 (2008)
13. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The Curious Case of Neural Text Degeneration. In: ICLR (2020)
14. Hotelling, H.: Relations between two sets of variates. In: Breakthroughs in statistics: methodology and distribution, pp. 162–190 (1992)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: ICML. pp. 4904–4916 (2021)
17. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual Prompt Tuning. In: ECCV. pp. 709–727 (2022)
18. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137 (2015)
19. Kaur, P., Pannu, H.S., Malhi, A.K.: Comparative Analysis on Cross-Modal Information Retrieval: A Review. Computer Science Review **39**, 100336 (2021)
20. Kim, S., Xiao, R., Georgescu, M.I., Alaniz, S., Akata, Z.: Cosmos: Cross-modality self-distillation for vision language pre-training. In: CVPR. pp. 14690–14700 (2025)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
22. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
23. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked Cross Attention for Image-Text Matching. In: ECCV. pp. 212–228 (2018)
24. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: ICML. pp. 19730–19742 (2023)
25. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: ICML. pp. 12888–12900 (2022)
26. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In: NeurIPS. pp. 9694–9705 (2021)

27. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
28. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: CVPR. pp. 23390–23400 (2023)
29. Li, Z., Hoiem, D.: Learning without forgetting. IEEE TPAMI **40**(12), 2935–2947 (2017)
30. Liao, L., Ma, Y., He, X., Hong, R., Chua, T.S.: Knowledge-Aware Multimodal Dialogue Systems. In: ACM MM. pp. 801–809 (2018)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
32. Liu, W., Zhu, F., Wei, L., Tian, Q.: C-CLIP: Multimodal continual learning for vision-language model. In: ICLR (2025)
33. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. NeurIPS **30** (2017)
34. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. ICLR (2017)
35. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS **32** (2019)
36. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
37. Ni, Z., Wei, L., Tang, S., Zhuang, Y., Tian, Q.: Continual vision-language representation learning with off-diagonal information. In: ICML. pp. 26129–26149 (2023)
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: NeurIPS. pp. 8024–8035 (2019)
39. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
42. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge (2008)
43. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioningwith reading comprehension. In: ECCV (2020)
44. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
45. Tan, H., Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In: EMNLP. pp. 5100–5111 (2019)
46. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2:

Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
47. Wang, K., Herranz, L., van de Weijer, J.: Continual Learning in Cross-Modal Retrieval. CVPRW (2021)
48. Weber, E., Marzo, N., Papadopoulos, D.P., Biswas, A., Lapedriza, A., Ofli, F., Imran, M., Torralba, A.: Detecting natural disasters, damage, and incidents in the wild. In: ECCV (August 2020)
49. Weber, E., Papadopoulos, D.P., Lapedriza, A., Ofli, F., Imran, M., Torralba, A.: Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents. IEEE TPAMI **45**(4), 4768–4781 (2022)
50. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
51. Yamaguchi, O., Fukui, K., Maeda, K.i.: Face recognition using temporal image sequence. In: IEEE International Conference on Automatic Face and Gesture Recognition. pp. 318–323 (1998)
52. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)
53. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. TMLR (2022)
54. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
55. Zhang, H., Yang, Y., Qi, F., Qian, S., Xu, C.: C2mr: Continual cross-modal retrieval for streaming multi-modal data. In: ACM MM. pp. 8963–8974 (2023)