One Patch to Caption Them All: A Unified Zero-Shot Captioning Framework

Lorenzo Bianchi[1,2,*] Giacomo Pacini[1,2,*] Fabio Carrara[1] Nicola Messina[1]
Giuseppe Amato[1] Fabrizio Falchi[1]

[1]ISTI-CNR, Italy [2]University of Pisa, Italy [*]Equal contribution

Abstract

Zero-shot captioners are recently proposed models that utilize common-space vision-language representations to caption images without relying on paired image-text data. To caption an image, they proceed by textually decoding a text-aligned image feature, but they limit their scope to global representations and whole-image captions. We present a unified framework for zero-shot captioning that shifts from an image-centric to a patch-centric paradigm, enabling the captioning of arbitrary regions without the need of region-level supervision. Instead of relying on global image representations, we treat individual patches as atomic captioning units and aggregate them to describe arbitrary regions, from single patches to non-contiguous areas and entire images. We analyze the key ingredients that enable current latent captioners to work in our novel proposed framework. Experiments demonstrate that backbones producing meaningful, dense visual features, such as DINO, are key to achieving state-of-the-art performance in multiple region-based captioning tasks. Compared to other baselines and state-of-the-art competitors, our models achieve better performance on zero-shot dense captioning and region-set captioning. We also introduce a new trace captioning task that further demonstrates the effectiveness of patch-wise semantic representations for flexible caption generation. Data and code are available at https://paciosoft.com/Patch-ioner/.

1. Introduction

Image captioning is one of the most representative tasks in vision-language understanding and has reached outstanding accuracy thanks to the availability of pre-trained vision-language backbones and large paired image-text datasets. In its basic formulation, a captioning model takes a full image as input and autonomously decides which elements must be described and up to what degree. To enable user guidance and produce more targeted descriptions, some previous works proposed region-level captioning methods [10, 23], which take as an additional input a spatial indication (e.g., bounding boxes) specifying which image regions have to be described and, possibly, in which order.

These region-level captioning methods require expensive manually labeled data to fully supervise the model. Indeed, each sequence or set of bounding boxes for a given image should correspond to a manually written ground-truth caption describing those objects. This fully supervised solution does not scale properly.

In this paper, we propose a perspective shift that enables us to perform region-level captioning with arbitrary spatial granularity — from a single image patch up to the entire image — *without requiring any form of region-level supervision or paired text-region data*. Specifically, instead of sticking with the classic setup where the subject of a captioning method is the *image* — then potentially conditioned on a set of sub-regions — we instead build on two straightforward yet powerful ideas: i) the simplest element that we could caption is a *patch*, the atomic element of an image representation in modern architectures based on vision transformers [12], and ii) we can easily aggregate multiple patch representations from frozen vision-language backbones to produce descriptions for arbitrarily large — and also potentially not contiguous — image regions.

We present a regional captioning framework which implements these ideas through a zero-shot regional captioning setup: we build on the assumptions that paired text-region data is not available at training time, and the models can be solely trained on text data, assumed to be available in large quantities. Our formulation offers maximum flexibility in zero-shot regional captioning tasks, producing models that effortlessly generate captions for various aggregations of image patches, ranging from individual patches to larger image regions, up to providing a caption for the entire image.

Despite the powerful perspective change that defines the patch as the new captioning unit, the problem is now entangled in a simple yet critical question: *how can we craft a model able to provide patch-level captions without relying on any direct patch-level ground truth supervision?*

In the last years, vision-language foundation models like CLIP [22, 32, 51] solved many downstream tasks in zero-

shot or even training-free configurations. In particular, contrastively learned vision-language representations enabled impressive results in zero-shot settings in image classification [51, 69], open-vocabulary detection [39, 72], and segmentation [16, 35], or text-image retrieval [27]. Image captioning, however, cannot directly employ CLIP machinery at inference time to generate text, given that CLIP is inherently a discriminative — and not a generative — approach. Only recently, image captioning models became zero-shot by decoupling image encoding — where pre-trained vision-language models like CLIP are used to create proper image and text representations — from the actual generative module. This is the case for models like [15, 18, 33, 43, 56, 59, 63, 67, 68], which i) employ CLIP to leverage a shared vision-language semantic space, and ii) train a text decoder on solely text samples to recover the text back from the CLIP textual feature. This requires nothing more than a pre-trained contrastive model and a large set of sole text samples to craft a powerful captioner.

In this paper, we show that many zero-shot captioners, paired with the right components, can be easily restructured to perform zero-shot *region-based* captioning. Therefore, we identify and study in detail the most critical components of this novel zero-shot regional captioning framework. Particularly, we focus our attention on the pre-trained vision-language contrastive backbone, which should be able, unlike CLIP, to create meaningful patch representations. To this aim, we largely explore DINO-based [7, 45] variants, having better localized capabilities than CLIP. On top of this, we address multiple modality-gap mitigation strategies helping the text decoder to correctly interpret visual features without the need for paired image-text data, as well as a study on different patch aggregation methods.

By analyzing existing components and employing vision backbones able to output patch-level meaningful representations like DINO, we show that we can enable many zero-shot captioners to reach state-of-the-art or comparable results in many zero-shot captioning task variants requiring captioning sub-parts of the entire image — *dense captioning* [23], *region-set captioning* [10], up to the standard image captioning [59] where the region to caption extends over the entire image. To better showcase the effectiveness of our framework in extreme patch-based captioning scenarios, we also introduce the *trace captioning* task, requiring the captioning of an image region specified by a mouse trace.

To summarize, our contributions are the following: a) we reformulate captioning by shifting perspective from the *image-to-caption* approach to a *patch-to-caption* one, unifying local and global tasks in one framework which does not require region-level supervision, through the exploitation of frozen vision-language backbones, b) we repurpose existing models to work within this novel framework, by analyzing the role of the key components, with a special attention to the vision backbone, c) we show the performance of these models on four zero-shot captioning tasks spanning different region granularity, from captioning few patches to the whole image, showing the effectiveness of the proposed perspective shift proposed by our framework despite its simplicity.

2. Related Work

Language-aligned Dense Image Representations are crucial for our goal of captioning at patch level. Vision-language models (VLM) like CLIP [51] introduced a powerful approach to learning global modality representations in a shared space via contrastive learning, paving the way to solve several downstream tasks, including captioning [11, 40]. However, in zero-shot settings, CLIP-like representations are known to struggle with dense tasks due to misalignment between local visual patches and fine-grained semantics [6, 52, 71]. On the other hand, visual-only self-supervised models (SSM) like DINO [7, 45] excel in local semantic modeling but lack a bridge with language. Recent works like SILC [42], DINO.txt [24] and SigLIP 2 [60] attempt to get the best of both worlds by combining DINO- and CLIP-like training objectives, aiming to obtain language-aligned dense representations. INViTE [8] modifies CLIP's visual encoder by zeroing the attention weights from each patch to all the others in the last layers, leading to more semantic patch features. DenseCLIP [53] and RegionCLIP [71] extend CLIP with additional region-level supervision. RegionCLIP leverages a region-proposal network to construct region-text pairs from image-text datasets, while DenseCLIP introduces a pixel-to-text matching loss to strengthen the alignment between local regions and textual concepts. Other methods instead exploit already existing VLMs and SSMs to get the same properties with minimal or no training: Talk2DINO [5] connects language to the DINOv2 space by mapping CLIP textual representations to DINOv2 patches. ProxyCLIP [29] instead leverages DINO's attention maps to improve the local properties of the CLIP visual embeddings of patches.

Zero-shot Image Captioning methods rely mostly on global CLIP representations to guide text generation. *Early-guided* decoding methods take CLIP visual features as input and introduce adaptation techniques to reduce the visual-textual modality gap [36]. DeCap [33] projects CLIP visual features into a more text-aligned space using a memory of texts as basis, while CapDec [43] and CLOSE [18] inject noise during text-only training to enable decoding also from the CLIP visual space. Diffusion Bridge [30] further mitigates this gap by applying a diffusion model to refine visual features toward the textual manifold. To improve the generation ViECap [15], MeaCap [68], MERCap [67] and EntroCap [63] leverage *external knowledge* to condition the decoding together with the CLIP image representation. *Late-guided* decoding methods instead use CLIP as a scoring or

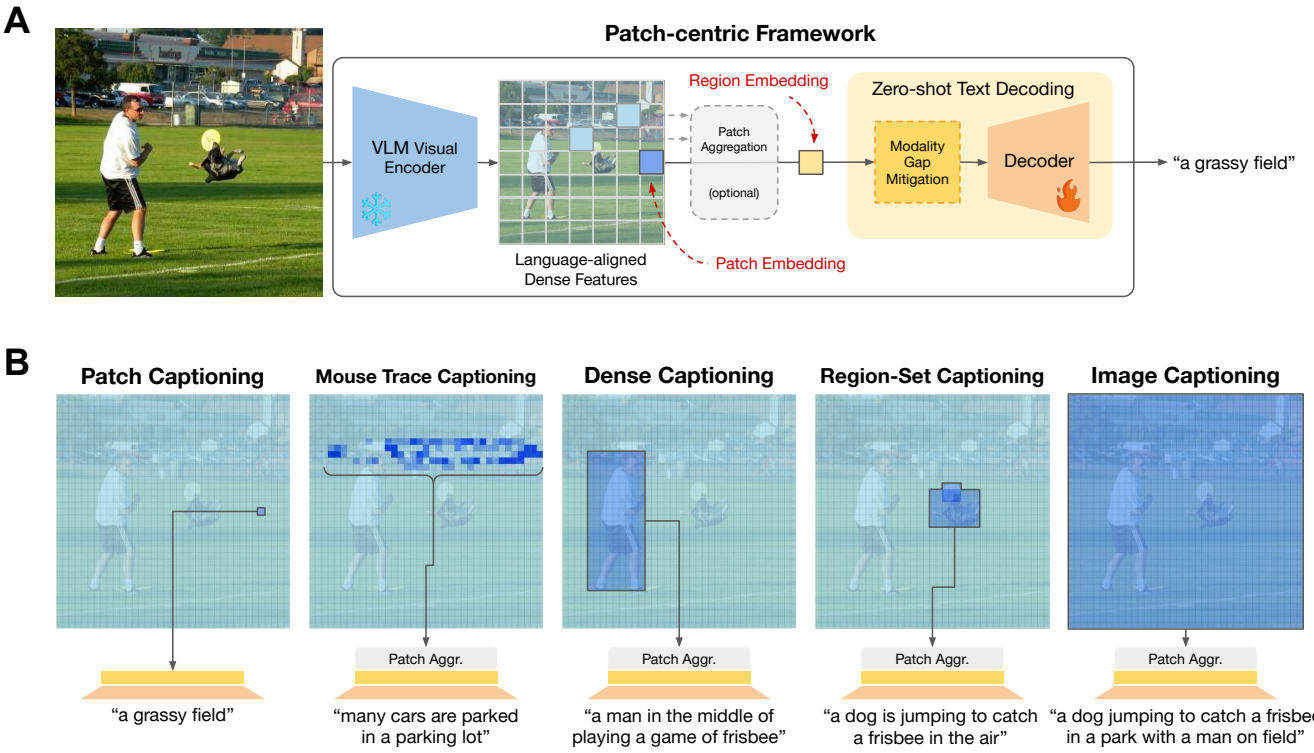

Figure 1. **Patch-centric framework for unified zero-shot captioning**. **A**. Overview of our framework. First, we extract language-aligned dense patch embeddings from the image using a VLM. Given a region, we select the underlying patches and aggregate their features to obtain a region representation. Finally, we obtain the region caption by applying a zero-shot text decoder, that is a) conditioned on the latent region representation, b) trained on text-only data, and c) equipped with a mechanism to handle the modality gap present in vision-language common spaces. This enables regional captioning without requiring region-level supervision. **B**. By aggregating patch-level features from arbitrary image regions, we can flexibly handle multiple captioning tasks across spatial granularities in a unique model.

optimization signal rather than direct input. ZeroCap [59] leverages CLIP gradients to steer the cached context during text generation, while MAGIC [56] optimizes token selection based on CLIP similarity scores. However, all the above approaches rely on global representations, which are not well-suited for capturing localized semantic details, making them suboptimal for patch-level or region-level captioning in zero-shot settings.

Region-level Captioning comprises several tasks in which models are asked to produce natural language descriptions based on sub-parts of an image. They pose additional challenges as naively captioning the cropped regions or feature maps often induces a loss of the global context of the image and, thus, misinterpretation of the region. For this reason, zero-shot solutions to this family of problems are still underexplored. For *controllable captioning* [10] — the generation of an image caption controlled by a set or sequence of regions — and *dense captioning* [23] — the localization and captioning of salient regions of an image — state-of-the-art solutions like CAG-Net [65], GRiT [62], ControlCap [70], and FlexCap [14] provide good performance but need supervision with ground-truth boxes. Recent works [19], [21] moved towards the direction of arbitrary regions captioning exploiting region-level supervision, and the Localized Narratives dataset [48] — comprising images, timed captions, and timed mouse tracks — provide the ingredients for evaluating captioning also at track- or patch-level. We propose a unique framework to tackle captioning at various granularities, from image- to patch-level, in a *zero-shot setting*.

3. A Patch-centric Framework

Overview. Figure 1 provides an overview of our framework, which reformulates captioning around image patches as the fundamental visual units. An input image is first encoded by a frozen vision–language backbone into dense, language-aligned patch embeddings. A region is then represented by aggregating the embeddings of its constituent patches, and a text decoder generates a natural-language caption from this aggregated representation.

Assumptions. Our approach operates under three main assumptions: (i) no region–text supervision is available during training, meaning the model never observes captions paired

with localized image regions, (ii) the vision–language backbone remains frozen, providing patch-level representations that reside in, or can be projected into, a joint embedding space shared with the text encoder, and (iii) no image-text supervision is available to train the text decoder, which instead can be trained exclusively on text data, learning to reconstruct captions from their textual embeddings.

At inference, the decoder interprets visual features through the shared multimodal space, enabling region-level captioning without any region–text or image-text annotations. The following subsections detail how each component is instantiated, how regions are represented through a parameter-free patch aggregation, and how we address the modality gap between visual and textual representations.

3.1. Motivation and Formulation

Learned region fusion requires regional data. Traditional regional captioning [14, 23, 70] follows an early injection of region specification in the model. Formally, given an image I and a region R, the region caption t is modeled as $t = \mathcal{D}(I, R)$. Such models require region-caption annotations and often use dedicated models or losses per captioning task or granularity. A step forward can be moved by introducing a formulation which disentangles image encoding and postpones region specification, defining $t = \mathcal{D}(\psi(I), R)$, where $\psi(I)$ provides a visual representation of the image independent of the region R, and $\mathcal{D}$ performs late region selection and text decoding. In order to avoid training the whole pipeline using region-level labels, we further decompose the decoder module $\mathcal{D}(\cdot)$ into two distinct modules: a parameter-free fixed aggregation agg_R of patch-level representations, and an actual text decoder $\phi(\cdot)$ not directly conditioned on regions R, so that $t = \phi(\mathrm{agg}_R(\psi(I), R))$. We will detail these two factorized components in the following paragraphs.

Parameter-free patch aggregation. Let $I \in \mathbb{R}^{H\times W\times 3}$ be an image split into non-overlapping patches of size $P \times P$. Each patch is encoded with a vision backbone ψ_v, yielding a dense grid of patch-level embeddings $V = \psi_v(I) = \{\mathbf{v}_i\} \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$, where $\mathbf{v}_i \in \mathbb{R}^D$. Assuming that $\psi_v(I)$ extracts this spatial grid of patch embeddings, a region definition R selects a subset of patch features that we aggregate to obtain the region embedding $\mathbf{v}_R = \mathrm{agg}_R(\psi_v(I))$. We describe this aggregation as $\mathbf{v}_S = \sum_{i\in S} w_i \mathbf{v}_i$, where S is the set of indices of patches that underlie the region, and w_i are aggregation weights. While several aggregation functions can be considered (e.g., uniform, gaussian, attention-based), we found that the specific choice has limited impact in regional captioning (see SM§11) and report results with mean aggregation ($w_i = 1/|S|$). Using a set operator as aggregation gives us the flexibility to aggregate arbitrary sets of patches, and thus, define regions as boxes, masks, traces, single patches, or full-image grids. Empirically, we also find that not all ψ_v are suitable to extracting patch-level meaningful visual representations; transformer architectures pretrained with dense local contrastive objectives (such as DINO) are significantly more robust at the patch level, as validated in §4.2.

Zero-shot decoding. We train a text decoder $\phi : \mathbb{R}^D \to \mathcal{T}$ using a prefix language modeling approach, where the decoder reconstructs a caption $t \in \mathcal{T}$ from its text embedding $\psi_t(t)$. If $\psi_t(t)$ is aligned to the visual encoder $\psi_v(I)$, we could directly decode the visual embeddings using a text-only decoder trained on text-only data. In such a case, we can finally decode the region representation into a natural language caption $t = \phi(\mathbf{v}_R)$. However, the assumption that $\phi(\cdot)$ can digest features from $\psi_v(I)$ while being trained to reconstruct features from $\psi_t(t)$ is often too optimistic due to the prominent modality gap [36]. In fact, text and image representations, despite being in the same multimodal space, occupy different, separated subspaces. To obtain a decoder applicable to visual embeddings, we experimented with three mitigation strategies for this gap. The first strategy, following [33], introduces a projection step at inference that maps visual features into the text subspace using a memory of text embeddings. The second, inspired by [18] and [43], trains the decoder under input perturbations, enforcing robustness to visual embeddings laying in another subspace. Finally, following [30], we explore a diffusion-based strategy that applies a diffusion model trained on the textual manifold to the patch-level representation, effectively "bridging" the gap by iteratively refining the visual features toward the text subspace. We analyze the effects of the modality gap mitigation strategies in SM§12.

Overall, our formulation offers three main advantages: a) it is zero-shot by design, in the sense that it only requires textual descriptions to be trained and does not require paired image-text samples to train the text decoder, b) any region (e.g., whole image, box, mask, free-form trace, single point) is addressed identically, supporting a modular, general-purpose regional captioning pipeline, and c) our method requires only a single forward pass of the vision backbone to extract patch features for an entire image, which can then be reused to caption multiple regions without rerunning the full pipeline for each one.

3.2. From Patches to Regions

Building on patch-level zero-shot captioning, our framework can generate captions for arbitrary regions of an image. A region is defined as a set of patches, and its representation is obtained by averaging the embeddings of its constituent patches. This simple formulation unifies several existing region-level captioning tasks, which differ only in how the relevant patches are selected, allowing us to address them without task-specific modifications. In the following, we

describe the tasks considered in our evaluation and their induced patch selections.

Image Captioning involves generating a single caption that describes the entire image. To achieve this, we derive a global representation $\mathbf{v}_I = \text{avg}_i(\mathbf{v}_i)$ by aggregating the feature embeddings of all patches $\{\mathbf{v}_i\}$ within the image I.

Dense Captioning requires locating salient regions in an image and generating their descriptions. Following defined evaluation protocols [23], we focus on captioning already defined boxes, effectively removing the localization subtask, which can be tackled using additional region-proposal models. Given a bounding box B and the set S_B of indexes of patches that intersect with B, we obtain the representation of the region $\mathbf{v}_B = \text{avg}_{i \in S_B}(\mathbf{v}_i)$.

Region-set Captioning consists of generating a single caption for multiple regions within an image, where each region is specified by a distinct bounding box. Given an image I and a set of bounding boxes $\mathfrak{B} = \{B_1, B_2, \ldots, B_K\}$, we define S_{B_k} as the set of patches that intersect with the k-th bounding box in $\mathfrak{B}$. To represent the entire set of regions, we aggregate the feature embeddings from all selected patches across all bounding boxes, which results in a combined region-level representation

$$\mathbf{v}_{\mathfrak{B}} = \frac{1}{\sum_{B \in \mathfrak{B}} |S_B|} \sum_{B \in \mathfrak{B}} \sum_{i \in S_B} \mathbf{v}_i \,. \tag{1}$$

Note that if a patch appears in more than one box, it is weighted more in the average.

Trace Captioning. To demonstrate the flexibility of our approach, we introduce *Trace Captioning*, a novel task in which the region of interest is specified by a mouse trace $T = \{p_1, \ldots, p_L\}$ with L points. Each point p_j is mapped to the corresponding patch index i_j, yielding the sequence $S_T = [i_1, \ldots, i_L]$. The trace-level representation is then obtained by averaging the embeddings of the selected patches

$$\mathbf{v}_T = \frac{1}{L} \sum_{j=1}^{L} \mathbf{v}_{i_j} \,. \tag{2}$$

Unlike box-based settings, this formulation allows for free-form, user-specified regions and thus enables interactive and fine-grained localized descriptions, expanding the scope of region-level captioning.

4. Experiments

In this section, we first quantitatively assess the role of the backbone in our novel framework (§4.2). Then, we measure the performance of our new formulation over state-of-the-art zero-shot captioning methods, evaluating them region-specific zero-shot tasks (§4.3). Other interesting yet less influential studies — like the role of the modality-gap mitigation strategy and the patch aggregation operator $\text{agg}_R(\cdot)$ — are available in SM §11 and §12.

To create a common ground, we first introduce the employed datasets and metrics in the following section.

4.1. Datasets and Metrics

Trace Captioning. We build a benchmark for Trace Captioning exploiting Localized Narratives [49] — a dataset in which annotators vocally described objects in images while moving the mouse pointer over the described object. The dataset provides temporal annotated voice transcriptions and mouse traces for the images of many standard captioning datasets. We took the labeled COCO [9, 38] with splits defined in [26] and Flickr30K[66] test subsets to build the trace captioning evaluation datasets. We split long traces and transcriptions for each image into sentences, and we discard the parts of the traces that are not temporally located between the start and the end of each sentence. We discarded noisy sentences — such as the ones describing image properties (*The image is blurred*, *the image is edited*, ...) and rewrote each sentence removing uncertainties typical of voice descriptions in a more concise and caption-like style through a few-shots prompted LLM (LLama 3 [13]). After annotation cleaning, 51 COCO images resulted without clean sentences and were discarded. Each sub-trace and relative sentence comprise an independent sample, that is, we ignore the temporal information of consecutive sentences and focus on evaluating the description of each sub-trace only. The final benchmark includes 4,949 COCO images with 14,283 captions (average caption length: 7.8 words) and 1,000 Flickr30K images with 26,614 captions (average caption length: 6.7 words). Samples and additional details are available in §13.

Dense Captioning. We assess the performance on dense captioning tasks following the evaluation procedure of [23], omitting the bounding box proposal and evaluating only the bounding box captioning task, using ground-truth boxes as input for the models. In addition to standard caption metrics, for this task, we also report the mAP as originally defined by [23]. We use the Visual Genome (VG) v1.2 [23, 28] and VG-COCO test splits [34]. The former comprises 5000 images from VG, while the latter contains 2476 images present in both VG and COCO. Both contain multiple bounding box annotations per image with descriptions.

Region-Set Captioning We follow the evaluation protocol of [10] that originally introduced region-set captioning. We use the Flickr30K Entities [47] and the COCO Entities [10] datasets. Each record comprises an image, a set of bounding boxes of variable length, and a ground-truth controlled caption. We evaluate on the test splits, comprising images from

the Karpathy [26] splits, which consist of 3569 and 1000 images for COCO and Flickr30k versions, respectively.

Image Captioning We follow the standard evaluation for zero-shot captioners, generating captions for the 5000 images in Karpathy's COCO test split. We compare with DeCap [33], CLOSE [18], ZeroCap [59], MAGIC [56], ViECap [15], CapDec [43], EntroCap [63], MeaCap[1] [68], MERCap [67], IFCap [31], and Diffusion Bridge [30].

Metrics. All datasets used in our evaluation provide a ground-truth caption for each annotation. We therefore adopt standard captioning metrics to assess the similarity between generated and reference captions. In particular, we focus on CIDEr (C) [61], which captures syntactic overlap, and RefPAC Score (P) [54], a more recent metric that quantifies semantic similarity independently of caption phrasing. For completeness, results with other traditionally used metrics — BLEU@4 [46], METEOR [4], ROUGE-L [37], and SPICE [2] — are reported in supplementary materials, as they follow the same trends. For the image captioning task, we additionally report CLIP-Score [20], which measures the alignment between an image and its generated caption in the joint CLIP space. This metric is not applicable to region-set, trace, or dense captioning tasks, where captions describe local regions rather than the entire image.

4.2. Backbone Selection

The choice of the visual backbone is crucial for our patch-centric framework, as the quality and semantic richness of the patch features directly impact the captioning performance. We tested several state-of-the-art vision-language models pre-trained *without region-level supervision* and evaluated their effectiveness within the framework. We tested vanilla **CLIP** [51], three CLIP adaptations for dense tasks — **DenseCLIP** [53], **INViTE** [8], and **ProxyCLIP** [29] —, **SigLIP2** [60], and two methods with visual encoders based on DINOv2 [44] — **DINO.txt** [25] and **Talk2DINO** [5]. In §7 we provide further details on each backbone adopted.

Patches are aggregated as described in §3.2, while for the zero-shot decoder, we align with the setting of [33], and use a prefix GPT-2 style decoder (SM§6 reports implementation details), with a memory-based latent projection approach as mitigation strategy for handling the modality gap. Specifically, before decoding, the region representation $\mathbf{v}$ is projected into the text embedding space as a similarity-weighted linear combination of memory elements, $\mathbf{v}_{\text{proj}} = M\,\alpha$ with $\alpha = \text{softmax}\big(\frac{1}{\tau} M^\top \mathbf{v}\big)$, where $M = [\mathbf{m}_1, \ldots, \mathbf{m}_N]$ stores the text embeddings $\mathbf{m}_j = \psi_t(t_j)$ and $\tau > 0$ controls the sharpness of the weighting distribution. This choice enables us to be directly comparable with other works using the same architecture in the all the following experiments, besides also performing the best among the tested zero-shot decoder methods (see SM§12). In SM§11, we also study how the choice of the memory bank M affects the captioning performance, providing an upper bound to metrics.

Table 1 shows that backbone effectiveness in our framework is closely tied to capturing fine-grained local semantics. Standard CLIP performs poorly, indicating that its patch tokens lack the spatial detail needed for our tasks [6, 41, 52]. Backbones that strengthen CLIP local representations, such as INViTE and DenseCLIP, achieve stronger results, supporting our hypothesis. The best performance comes from DINOv2-based models, including DINO.txt and Talk2DINO, with the latter emerging as the most effective encoder. This underscores the importance of semantically rich patch-level features for high-quality region-level captions. For this reason, we show results using Talk2DINO as the default backbone in subsequent experiments.

4.3. Comparison with SOTA

Despite significant advances in zero-shot image captioning, we are unaware of any prior methods specifically tailored for zero-shot regional captioning tasks. Existing zero-shot captioners are usually evaluated only at the level of whole images, without any mechanisms to natively attend to arbitrary regions. To rigorously quantify the benefit of our approach, in Table 2, we compare against both state-of-the-art zero-shot image captioners and adapted baselines: (i) *whole-image zero-shot captioners* in their standard setting or applied to region crops, simulating regional captioning by isolating local content, and (ii) *region-supervised encoders*, thus outside our no-region-label setting, that leverage mask-based (AlphaCLIP, [57]) or crop-based (RegionCLIP, [71]) attention coupled with the same zero-shot decoder, allowing them to attend to specific regions. This design ensures that our evaluation covers the strongest available baselines for both image-level and region-level zero-shot captioning. While we exclude models trained on region-level data [19, 21], in §8 we compare our framework against representative LMMs. Despite their massive end-to-end pre-training on image-text pairs, our text-only method achieves superior performance on fine-grained regional tasks.

In addition to our strongest model (i.e., Talk2DINO with the memory-based mitigation strategy), we also report other combinations that express existing zero-shot captioning approaches (CLOSE [18], CapDec [43], VieCap [15], MeaCap [68] and Diffusion Bridge [30]) but replacing the original CLIP backbone. Specifically, CLOSE and CapDec can be expressed by choosing the noise-injection mitigation strategy and using the standard decoder pipeline described in §4.2. The same applies to ViECap and MeaCap, although with the addition of extra knowledge in the text decoding

[1]We picked the MeaCap$_{InvLM}$, which uses a GPT-2 decoder and achieves the highest scores in the zero-shot captioning task.

Captioning Task: (Dataset)			Trace (COCO)		Dense (VG v1.2)		Region-Set (COCO Entities)		Image (COCO)		
	Vision Backbone	Text Backbone	C	P	C	P	C	P	C	P	CLIP-S
CLIP	CLIP B/16	CLIP B/16	10.9	75.0	10.9	74.2	41.6	78.8	42.1	84.0	66.2
DenseCLIP	CLIP B/16	CLIP B/16	18.6	75.3	19.9	75.2	51.0	77.6	28.0	77.0	57.3
INViTE	CLIP B/16	CLIP B/16	13.8	76.4	16.8	77.3	43.3	78.9	21.3	79.1	60.6
ProxyCLIP	DINO B/8 + CLIP B/16	CLIP B/16	16.7	75.7	15.7	76.0	41.2	78.4	28.7	79.0	61.7
ProxyCLIP	DINOv2 B/14 + CLIP B/16	CLIP B/16	16.5	75.7	15.5	76.0	40.6	78.5	27.4	78.6	61.0
SigLIP2	SigLIP2 B/16	SigLIP2 B/16	18.3	73.6	19.8	73.4	47.2	76.7	27.7	77.0	56.4
DINO.txt	DINOv2 B/14	DINO.txt	23.2	**78.8**	23.4	78.5	91.8	86.3	67.8	87.2	70.8
Talk2DINO	DINOv2 B/14	Talk2DINO	**27.9**	78.7	**31.9**	**78.8**	**109.1**	**87.5**	**69.2**	**87.4**	**72.8**

Table 1. **Vision-Language Backbones.** CIDEr (C) and RefPAC-S (P) across four captioning tasks.

step: ViECap uses external entity-aware prompts, while MeaCap leverages also structured concept retrieval from a knowledge base. Finally, Diffusion Bridge is implemented by applying a diffusion model trained on the textual manifold to the patch-level representation.

We use two datasets for each task — a COCO-derived dataset and an additional dataset such as Visual Genome [28] or Flickr30k [66]. Figure 7 shows qualitative results.

Patch-centric captioning excels in local, fine-grained tasks. In trace and dense captioning tasks, which emphasize local visual content, our patch-centric framework significantly outperforms all baselines across metrics. In trace captioning (Table 2, 1st group), our formulation outperforms image-based captioners and crop-based adaptations. Models relying on global CLS representations fail to capture the precise objects and attributes under the trace. Even AlphaCLIP, a region-supervised backbone applicable to traces, lags behind, underlying intrinsic limitations of the pretrained CLIP backbone. Dense captioning shows a similar trend (Table 2, 2nd group), with our models outperforming baselines. For this task, we report crop-based adaptations for DeCap, ViECap, MeaCap, IFCap, and Diffusion Bridge, as they provide a stronger baseline with respect to the same models applied to the global CLS (see Table 7 in SM). Although isolating regional content, these models discard broader contextual cues that are crucial for coherent dense descriptions.

Patch aggregation extends seamlessy to context-aware captioning. Also in region-set captioning (Table 2, 3rd group), our models achieves state-of-the-art results, outperforming both zero-shot baselines and region-supervised models. Region-set captioning tends to align more closely with image-level captioning rather than strictly focusing on localized regions (see Figure 7 and Figure 8 in SM), as regions are intended to control an image-level caption[2]. Thus, global models also tend to perform well on those tasks, showing a narrower gap with respect to ours. By aggregating patch embeddings from arbitrary and possibly disjoint sets of regions, the model produces coherent and contextually rich captions that align with the collective semantics of the chosen areas. In contrast, whole-image methods cannot naturally incorporate regional cues, limiting their effectiveness in this setting. Importantly, our patch-based aggregation even surpasses AlphaCLIP, which relies on explicit mask supervision.

Patch-centric models deliver comparable performance on whole-image captioning. In whole-image captioning (Table 2, 4th group), our results remain competitive with the strongest zero-shot captioners but are slightly behind *dedicated* image-centric architectures such as MERCap [67] and EntroCap [63]. For this task, we report the performance of our models using an attention-based weighting, when helpful, as it usually performs marginally better than the standard average patch aggregation and thus represents the best available model for this task and reaches the smaller gap with state-of-the-art models (see SM§11). Notably, adding structured external knowledge or filtered retrieval (as in ViECap or MeaCap) on noise-based decoders improves fluency and informativeness, suggesting such modules are complementary for strong regional semantics. However, they compare similarly to the model using the memory-based decoder.

5. Conclusions

We introduced a zero-shot framework that shifts from an image-centric to a patch-centric approach, enabling captioning for individual patches and arbitrary aggregations without any region or image supervision. We rely on the strong spatial awareness of DINOv2, whose local image patches have been bridged with the text modality, and we disentangle the training of the decoder network. This flexible and scalable method sets the new state of the art on various regional captioning tasks, including dense and region-based captioning, as well as our newly proposed trace captioning. Results show

[2]This is expected since the ground-truth captions in the COCO Entities dataset originate from the image-level annotations of COCO [10].

Model	Trace Captioning				Dense Captioning						Region-Set Captioning				Image Captioning					
	COCO		Flickr30k		VG v1.2			VG-COCO			COCO Entities		Flickr30k Entities		COCO			Flickr30k		
	C	P	C	P	mAP	C	P	mAP	C	P	C	P	C	P	C	P	CLIP-S	C	P	CLIP-S
Whole-image Zero-shot Captioners																				
ZeroCap [59] CVPR'22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14.6	-	-	-	-	-
MAGIC [56] ArXiv'22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.3	-	-	17.5	-	-
CLOSE [18] ICCV'23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	81.2	-	-	-	-	-
CapDec [43] EMNLP'22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	91.8	-	-	35.7	-	-
EntroCap [63] NEUCOM'25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.3	-	-	41.5	-	-
MERCap [67] AAAI'25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**96.0**	-	-	**45.6**	-	-
ViECap† [15] ICCV'23	24.3	74.4	12.0	68.8	14.90°	26.4°	74.3°	15.2°	26.6°	74.3°	102.7	85.0	31.8	74.9	89.7	88.5	75.6	29.8	80.6	70.5
MeaCap† [68] CVPR'24	22.5	74.4	12.6	69.8	15.01°	28.6°	75.1°	16.0°	28.9°	75.1°	97.9	85.2	38.6	76.4	86.0	88.6	77.8	40.1	82.8	73.9
DeCap† [33] ICLR'23	20.5	75.3	11.2	71.0	17.75°	24.6°	77.8°	17.8°	24.9°	77.7°	95.1	87.4	39.4	78.8	87.4	90.6	**79.3**	40.0	84.8	**76.3**
IFCap† [31] EMNLP'24	20.8	73.2	10.0	67.4	13.7°	21.3°	72.5°	13.4°	22.8°	73.0°	95.0	83.9	30.5	73.3	86.9	87.3	74.6	29.4	79.2	69.2
Diffusion Bridge† [30] CVPR'25	21.8	74.8	10.7	69.8	14.9°	20.0°	72.8°	15.2°	20.2°	72.8°	94.7	85.3	32.9	75.8	85.1	88.8	76.7	33.7	82.1	72.4
With Region-level Supervision																				
RegionCLIP [71] CVPR'22 + Mem. (≃ DeCap)	-	-	-	-	15.85	21.7	76.7	16.01	21.0	75.4	-	-	-	-	93.4	**91.2**	77.5	38.8	84.5	73.6
AlphaCLIP [57] CVPR'24 + Mem. (≃ DeCap)	21.3	75.4	11.8	71.0	14.63	19.1	73.9	14.82	19.4	73.8	95.1	87.4	39.5	78.8	89.7	91.1	78.2	40.9	**85.4**	75.0
Patch-ioner (Our Patch-based Framework)																				
T2D + Mem. (≃ DeCap)	27.9	78.7	18.8	77.0	**21.31**	**31.9**	78.8	**21.53**	**32.3**	**78.7**	109.1	**87.5**	**44.1**	**79.1**	88.5◇	90.2◇	76.0◇	39.3◇	84.2◇	71.8◇
T2D + Noise (≃ CLOSE, CapDec)	29.3	78.1	19.3	75.6	20.26	26.3	77.0	20.33	26.4	76.9	97.5	85.6	37.1	76.5	65.5	86.2	70.9	27.8	80.8	67.0
T2D + Noise + External knowledge (≃ ViECap)	28.2	78.2	18.5	76.2	18.43	30.3	77.8	18.43	30.7	77.7	109.3	86.7	37.8	77.8	88.5◇	89.2◇	73.7◇	34.1◇	82.8◇	69.9◇
T2D + Noise + Filtered knowledge (≃ MeaCap)	27.4	**78.8**	**20.3**	**77.3**	18.66	**31.9**	**78.9**	19.43	**32.3**	**78.7**	104.4	86.9	42.3	78.6	83.0◇	89.6◇	74.8◇	39.4◇	84.4◇	71.4◇
T2D + Diffusion (≃ Diffusion Bridge)	**29.4**	77.6	18.7	75.0	19.01	30.7	77.1	19.40	31.1	77.0	**114.4**	86.6	40.8	77.1	92.8◇	89.0◇	74.2◇	36.4◇	82.3◇	69.0◇

†: reproduced by us. ◇: Model applied to image crops. ◊: attention-based weighting.

Table 2. Comparison of our Patch-ioner framework, using Talk2DINO (**T2D**), with ZS methods on trace, dense, region-set, and image captioning tasks. Our approach consistently outperforms whole-image and region-supervised baselines in local, fine-grained captioning tasks, while achieving competitive results on whole-image captioning. The table reports CIDEr (C), RefPAC (P), mean average precision (mAP) for dense captioning, and CLIP-Score (CLIP-S) when applicable; best and second-best results are in **bold** and underlined, respectively.

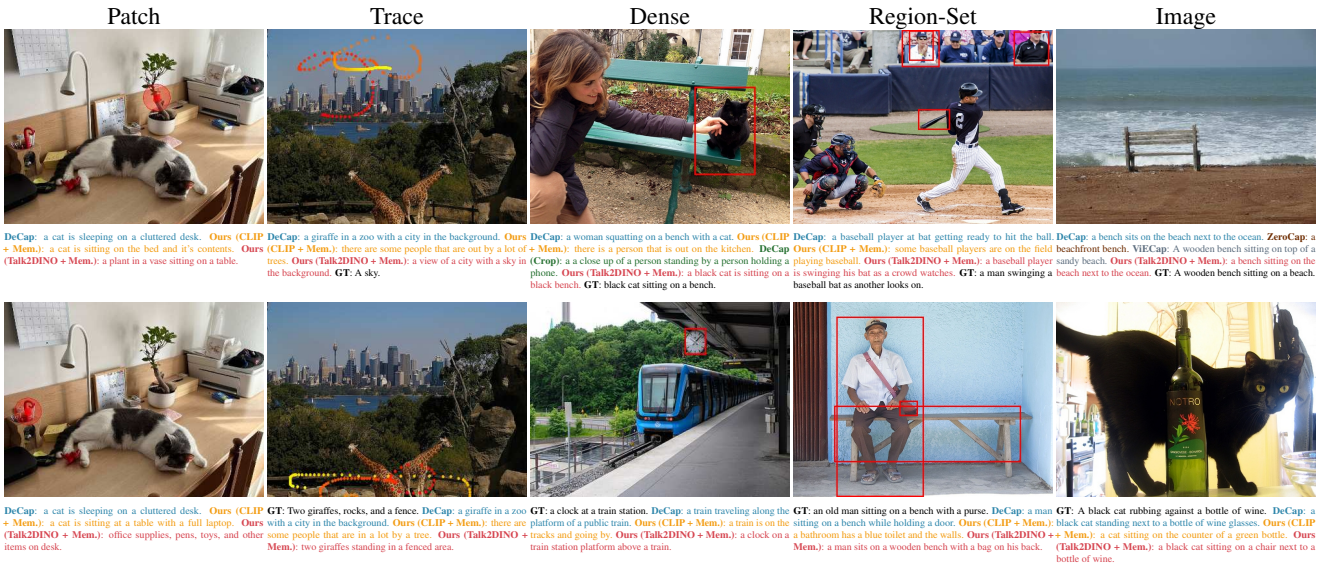

Figure 2. **Qualitative results** from finer (left) to coarser (right) tasks. Note the discrepancy of predicted and ground-truth captions when an image-level (DeCap, DeCap (Crop)) or a CLIP-based regional (CLIP + Mem.) captioner is applied, with respect to Talk2DINO-based model.

that our patch-centric approach can effectively bridge the gap between local and global understanding in image captioning, providing a unified framework for multi-granularity captioning tasks in a zero-shot setting. Moreover, our models require a single backbone forward pass to caption multiple regions, facilitating viability in interactive applications.

Limitations and Future Work. Despite strong zero-shot performance, our model still lags behind fully supervised, task-specific approaches. The contextual scope of each patch is fixed by the backbone and is not adjusted to meet the user's intents. Also, the modality gap introduces noise that can cause hallucinations. Future work will focus on improving patch-level semantics, e.g., through image-level captioning loss in weakly-supervised settings, or refine the patch-to-text projection to reduce the modality gap in zero-shot settings.

Acknowledgment

This work has been supported by the PNRRM4C2 project "FAIR - Future Artificial Intelligence Research", by the PRIN 2022-PNRR project "MUCES" (CUP E53D23016290001 and B53D23026090001), and by the PNRR project "ITSERR - Italian Strengthening of Esfri RI Resilience" (CUP B53C22001770006), all funded by the European Union - NextGenerationEU, and by the SUN XR project funded by the Horizon Europe Research & Innovation Programme (GA n. 101092612).

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 1

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 6

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[5] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. *arXiv preprint arXiv:2411.19331*, 2024. 2, 6, 1, 3, 5, 7

[6] Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. Improving fine-grained understanding in image-text pre-training. In *ICML*, 2024. 2, 6

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[8] Haozhe Chen, Junfeng Yang, Carl Vondrick, and Chengzhi Mao. Invite: Interpret and control vision-language models with text explanations. In *International Conference on Representation Learning*, pages 35589–35608, 2024. 2, 6, 1

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[10] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *CVPR*, 2019. 1, 2, 3, 5, 7, 11

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Generating more pertinent captions by leveraging semantics and style on multi-source datasets. *IJCV*, 132(5):1701–1720, 2024. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 1

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[14] Debidatta Dwibedi, Vidhi Jain, Jonathan J Tompson, Andrew Zisserman, and Yusuf Aytar. Flexcap: Describe anything in images in controllable detail. *NeurIPS*, 37:111172–111198, 2025. 3, 4

[15] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *ICCV*, pages 3136–3146, 2023. 2, 6, 8, 1, 5

[16] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 2

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[18] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language supervision. In *ICCV*, pages 2672–2683, 2023. 2, 4, 6, 8, 7, 12, 13

[19] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 3, 6

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6

[21] Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. In *CVPR*, pages 24763–24773, 2025. 3, 6

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1

[23] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016. 1, 2, 3, 4, 5

[24] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec,

Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. *arXiv preprint arXiv:2412.16334*, 2024. 2, 7

[25] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 6, 1

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 5, 6

[27] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Šuma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiří Matas, Ondřej Chum, and Giorgos Tolias. ILIAS: Instance-level image retrieval at scale. In *CVPR*, 2025. 2

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5, 7

[29] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation. In *ECCV*, 2024. 2, 6, 1

[30] Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4050–4059, 2025. 2, 4, 6, 8

[31] Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. *arXiv preprint arXiv:2409.18046*, 2024. 6, 8

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 1

[33] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 4, 6, 8, 1, 3, 5

[34] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8650–8657, 2019. 5

[35] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 2

[36] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35:17612–17625, 2022. 2, 4, 8

[37] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 6

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 5

[39] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 36:72983–73007, 2023. 2

[40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[41] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *CVPR*, pages 19413–19423, 2023. 6

[42] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *ECCV*, pages 38–55. Springer, 2024. 2

[43] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4055–4063. Association for Computational Linguistics, 2022. 2, 4, 6, 8, 3, 7

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 2

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5

[48] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language

with localized narratives. In *ECCV*, pages 647–664. Springer, 2020. 3, 9

[49] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, pages 647–664. Springer, 2020. 5, 11

[50] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1, 2, 6

[52] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual Grouping in Contrastive Vision-Language Models. In *ICCV*, 2023. 2, 6

[53] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 2, 6, 1

[54] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *CVPR*, pages 6914–6924, 2023. 6

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets & Benchmarks Track*, 2022. 3

[56] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 2, 3, 6, 8

[57] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *CVPR*, pages 13019–13029, 2024. 6, 8

[58] Gemma Team. Gemma 3. 2025. 3

[59] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, pages 17918–17928, 2022. 2, 3, 6, 8, 12, 13

[60] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 6, 1

[61] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[62] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *ECCV*, pages 207–224. Springer, 2024. 3

[63] Jie Yan, Yuxiang Xie, Shiwei Zou, Yingmei Wei, and Xidao Luan. Entrocap: Zero-shot image captioning with entropy-based retrieval. *Neurocomputing*, 611:128666, 2025. 2, 6, 7, 8

[64] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[65] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *CVPR*, pages 6241–6250, 2019. 3

[66] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 5, 7

[67] Delong Zeng, Ying Shen, Man Lin, Zihao Yi, and Jiarui Ouyang. Zero-shot image captioning with multi-type entity representations. In *AAAI*, pages 22308–22316, 2025. 2, 6, 7, 8

[68] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14100–14110, 2024. 2, 6, 8, 1, 3, 5

[69] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 2

[70] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. Controlcap: Controllable region-level captioning. In *ECCV*, pages 21–38. Springer, 2024. 3, 4

[71] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-Based Language-Image Pretraining. In *CVPR*, 2022. 2, 6, 8

[72] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes

using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 2

Supplementary Material

6. Implementation Details

For the training of the textual decoder of the memory-based configuration ϕ, we adopt a prefix GPT2-style decoder-only Transformer with 4 attention heads and 4 layers, following the architecture used by [33]. We train the model on captions from the COCO training set, which also serves as the memory bank M for the projection mechanism, comprising approximately 500k texts. We set the hyperparameters of the projection mechanism as in DeCap ($\tau = 0.01$), and use the AdamW optimizer with a weight decay of 0.01. Training proceeds for 10 epochs with a learning rate of 10^{-5} and a batch size of 64. A comprehensive overview of the framework operating in the DeCap setting (with the projection mechanism as modality gap mitigation strategy) is provided in Figure 4.

For the external knowledge-based captioning models we performed a training of 15 epochs with a batch size of 80 captions on the same GPT2-style textual decoder using a learning rate of 2×10^{-5} and a gaussian noise variance of 16×10^{-3} to replicate the experimental settings of [15] and [68].

All experiments were conducted on a single NVIDIA H100 GPU with 80GB of HBM3 memory. Training took approximately 25 minutes per epoch.

7. Backbone Details

We briefly summarize here the characteristics of the vision-language models we tested in our framework.

- **CLIP** [51]: A foundational model that learns a shared embedding space for images and text through contrastive learning. While being the most used model for global image-text alignment, its patch tokens are known to lack strong spatial and fine-grained semantic information. The input resolution of its training is 224 pixel.
- **DenseCLIP** [53]: A fine-tuned version of CLIP that incorporates a pixel-text matching loss to enhance the model's ability to understand local regions. The official implementation input resolution is 640 pixel for the ViT-B/16 version.
- **INViTE** [8]: This method modifies CLIP's vision transformer to bring patch tokens in the text space by disabling the self-attention mechanism. It employs the same visual encoder of CLIP, trained at 224 pixel input resolution.
- **ProxyCLIP** [29]: A model that leverages the local understanding of a DINO backbone to improve CLIP's patch-level representations. It achieves this by replacing the attention maps in CLIP's final layer with DINO's attention maps, effectively transferring DINO's fine-grained spatial awareness to the CLIP embedding space. The DINO ViT-B/8 version was tested with images at 296 pixel resolution, while the DINOv2 ViT-B/14 at 518 pixel.
- **SigLIP2** [60]: A multilingual vision–language model that improves upon CLIP by enhancing both global alignment and dense localization capabilities through refined training objectives and architectural adjustments. We employed the B/16 variant, which uses a ViT with 16-pixel patches and is trained at an input resolution of 512 pixels. To obtain text-aligned semantic patch representations, we follow the authors' dense feature extraction strategy and apply the MAP (Mean Attention Pooling) head—normally used to produce the global image representation embedding—individually to each patch token.
- **DINO.txt** [25]: This model builds upon a frozen DINOv2 backbone, adding learnable transformer blocks on top. It is then trained with a contrastive objective against a text encoder to align both global and patch-level representations with language. The DINOv2 backbone was trained at the resolution of 518 pixel.
- **Talk2DINO** [5]: This model creates a bridge between the CLIP and DINOv2 embedding spaces. It trains a projection to map CLIP text embeddings into the DINOv2 patch space, using DINOv2's highly meaningful attention maps to identify and align with the most relevant patches during training. The DINOv2 backbone was trained at the resolution of 518 pixel.

8. Comparison with LMMs

Our framework is designed as a zero-shot regional captioner, relying solely on text-only corpora to train the decoder, and operating without any paired image-text or region-text supervision. While most of the Large Multi-modal Models (LMMs) are trained without region-text supervision, they are trained on massive datasets using image-text pairs, making them inherently supervised solutions.

While this difference in training paradigm places LMMs outside the core assumptions of our zero-shot setup, we include a comparison with representative, high-performing LMMs — Llava1.5 OneVision (4B) [1], Qwen2.5 VL [3] (3B) and Qwen3 VL (4B) — to provide a strong, external supervised benchmark for our framework's capabilities.

8.1. LMM Adaptation for Regional Tasks

LMMs typically process a whole image and cannot natively process explicit regional coordinates (such as bounding boxes or traces) as additional inputs, unless specifically trained with region-level box or mask annotations. To enable a fair comparison of our regional tasks, we propose two zero-shot adaptation strategies to inject regional information into the LMM.

- **Visual Prompting:** The regional annotations (bounding boxes or traces) are drawn over the image, allowing the model to condition its generation on the spatial input. We then ask the model to describe the annotated image.

Captioning Task: (Dataset)	Adaptation strategy	Trace (COCO)		Dense (VG v1.2)		Region-Set (COCO Entities)		Image (COCO)		
		C	P	C	P	C	P	C	P	CLIP-S
LMMs										
Llava1.5 One-Vision (4B)	Visual Prompting	21.2	75.9	25.3	76.2	72.2	87.3	**91.6**	**91.1**	80.1
Qwen2.5 VL (3B)	Visual Prompting	17.4	74.0	19.1	74.2	59.6	84.5	77.9	89.9	81.2
Qwen3 VL (4B)	Visual Prompting	9.1	70.7	10.8	74.2	10.8	74.2	19.4	84.2	**85.3**
Llava1.5 One-Vision (4B)	Crop	19.7	75.4	15.1	72.7	75.9	86.7	**91.6**	**91.1**	80.1
Qwen2.5 VL (3B)	Crop	18.4	73.9	11.9	69.2	68.6	85.9	77.9	89.9	81.2
Qwen3 VL (4B)	Crop	7.6	69.3	3.5	68.8	20.2	80.4	19.4	84.2	**85.3**
Patch-ioner (Our Patch-based Framework)										
T2D + Mem. (0.21B)		**27.9**	**78.7**	**31.9**	**78.8**	**109.1**	**87.5**	69.2	87.4	72.8

Table 3. **Patch-ioner framework vs. LMMs.**

- **Cropping:** The image is cropped to the bounding box encompassing the regional annotation (be it a trace, a box, or a set of boxes). The cropped image is then fed to the LMM, forcing the model to focus its attention solely on the region of interest.

An example of adaptation for each task with the respective prompt is shown in Figure 3.

8.2. Results Analysis

Table 3 presents the comparison results of our Patch-ioner framework against state-of-the-art LMMs across various captioning granularities.

In the finer-level tasks, specifically Dense Captioning (VG v1.2) and the Trace Captioning (COCO), the pronounced performance gap between our framework and LMMs highlights the fundamental inability of LMMs to effectively reason at the precise region level based solely on a visual annotation or a simple crop without sufficient global context. For these granular tasks, the Visual Prompting strategy is generally preferable for LMMs because it preserves the essential context of the entire image. Furthermore, simple cropping is problematic for concave traces as it fails to precisely delineate the intended region of the image.

In tasks involving broader regions, such as Region-Set Captioning (COCO Entities), the Patch-ioner framework achieves better performance despite LMMs showing closer results, highlighting our higher capability to generate captions specifically guided by the set of input regions. For these broader regions, the Cropping adaptation strategy generally leads to better LMM performance than Visual Prompting, suggesting that Region-Set Captioning demands a lower level of context from the uncropped parts of the image with respect to Dense and Trace Captioning.

For the Image Captioning task, which relies on global understanding, LMMs generally demonstrate superior performance compared to our zero-shot captioner baseline. This is largely expected, as LMMs are bigger and extensively trained on massive image-text pairs.

8.3. Comparison Highlights

The comparison with powerful LMM baselines adapted for regional tasks clearly underscores the unique advantages of the Patch-ioner framework in terms of design, data requirements, parameter efficiency, and inference speed.

Designed for Granularity The fundamental difference lies in the design objective: our Patch-ioner framework is explicitly engineered for multi-granularity, region-level captioning , while LMMs are built primarily for global image understanding. The need for ad-hoc adaptation strategies for LMMs (Visual Prompting or Cropping) inherently limits their precision, resulting in significantly lower performance on fine-grained regional tasks –as shown in Table 3– compared to our method.

Data and Training Efficiency Unlike LMMs, which are trained on massive paired image-text datasets and are therefore inherently supervised solutions, our approach adheres to a strict zero-shot paradigm. The text decoder is trained solely on text corpora, eliminating the dependency on costly, large-scale image-text supervision for training the generative module.

Parameter and Computational Efficiency The Patch-ioner framework is dramatically more lightweight and parameter-efficient than LMMs. We provide a compact model (e.g., 0.21B parameters for the Talk2DINO configuration), contrasting sharply with the multi-billion parameter counts typical of LMMs (e.g., 3B to 4B parameters in the tested models).

Inference Speed and Scalability One of our framework's most critical advantages is its inference efficiency for multiple regions within a single image. For an image containing multiple annotations (such as in the Dense Captioning task, which can have over a hundred boxes), we require only a single forward pass of the frozen vision backbone to extract all patch features. These pre-computed patch features can then be reused to caption arbitrary regions. Conversely, an LMM adapted through Cropping or Visual Prompting requires a full inference of the vision-language model for every single annotation, leading to dramatically slower performance when scaling to dense or complex regional tasks

9. Ablations on Text Decoder Architectures

We tested different architectures and sizes for the text decoder network. In particular, we trained GPT-2 small [50], Gemma 3 270M [58], Qwen 3 0.6B [64], LLaMa 3.2 1B [17] and Qwen 3 1.7B [64]. Details on training hyperparameters are provided in §6. The dataset adopted for these trainings is made of the ground truth captions of COCO train Karpathy split. The total number of different textual captions for that split is 566,747. Table 4 reports the scores obtained when varying the decoder in our framework.

Across all tasks, we observe that increasing the capacity of the textual decoder does not translate into consistent improvements. GPT-2 small (124M) performs surprisingly strongly, outperforming larger models on Trace and Region-Set captioning in terms of CIDEr, while maintaining competitive RefPAC-S scores. Larger decoders such as Gemma 3 (270M), Qwen3 (0.6B), and Llama 3.2 (1B) yield similar or slightly worse results, and the largest tested model, Qwen3 1.7B, exhibits the lowest performance on all CIDEr metrics. These findings suggest that, under our training regime and dataset size (COCO train Karpathy split, 566k captions), model capacity is not the limiting factor. Larger decoders tend to overfit more easily and provide little benefit in a setting where the captioning supervision is narrow in distribution and relatively small compared to their pretraining scale. Moreover, since our overall architecture relies on a strong visual encoder capable of providing semantic local representations, the decoder's primary role is to map visual representations to fluent text; in this context, compact autoregressive models appear sufficiently expressive. Interestingly, RefPAC-S shows minimal variation across decoder sizes, indicating that semantic alignment with the image features (rather than surface-level text quality) is largely preserved even with larger models. However, the trend of decreasing CIDEr with increasing model size suggests that bigger decoders may introduce unnecessary linguistic diversity that hurts match-based metrics. Overall, these results highlight that changing the decoder affects only marginally the performances.

10. Ablations on Text-only Training Dataset

In this section, we compare the results obtained when using the classical training strategy adopted by the models such as [33, 43, 68] in the zero-shot captioning task — which consists of training for 10 epochs on the collection of texts from the COCO dataset — with training on a different dataset, to assess generalization capabilities of our framework.

For this comparison, we took the best Patch-ioner configuration — which consists of using Talk2DINO [5] as visual backbone and a memory bank of texts borrowed from COCO as in [33]— and we trained it on a subset of ReLaion.

10.1. Selected dataset: ReLaion 600M

ReLaion is a large-scale image-text dataset containing approximately 600 million image-caption pairs sourced from LAION-2B [55]. Each entry includes multiple machine-generated captions of varying quality, along with a "best caption" field obtained by ranking the alternatives using a CLIP-based scoring function. In our ablation we exclusively use this "best caption" attribute, since it provides a higher-quality textual signal while avoiding noisy or inconsistent descriptions that typically arise in large crawled datasets.

To ensure a fair comparison with the standard COCO-based training strategy from prior work [33, 43, 68], we adopt a subset of ReLaion sized so that one training epoch over that matches the number of training steps of the standard COCO training lasting 10 epochs. Since the COCO training split contains roughly 560k captions, one epoch on a 5.6M-sample subset of ReLaion results in approximately the same number of optimization steps, while a 28.3M-sample subset corresponds to five times more steps. In all experiments, we maintain the same Patch-ioner configuration (our best-performing model) and apply identical optimization settings.

This setup allows us to assess the generalization capability of our framework when trained on broader, noisier web-scale text compared to the highly curated COCO captions that are commonly used in zero-shot captioning pipelines.

10.2. Discussion

The results in Table 5 highlight three main trends.

(1) ReLaion training improves cross-dataset generalization. Training on ReLaion — even for a single epoch — consistently improves performance across captioning tasks compared to the classical COCO-only training. The 5.6M-sample setting yields clear gains on Trace and Dense Captioning, and the larger 28.3M subset further enhances Region-Set and Image Captioning scores. These improvements indicate

Figure 3. **LMM Adaptation Strategies for Zero-Shot Regional Captioning.** This figure visualizes the two input-adaptation strategies used to benchmark Large Multimodal Models (LMMs) on Region-level Captioning tasks: Trace, Dense, and Region-Set Captioning. **Top Row:** The original image corresponding to each task. **Middle Row (Visual Prompting):** The task region is spatially indicated by superimposed red annotations (traces or bounding boxes). The LMM input includes this visually prompted image along with a detailed instruction (shown below the images) to describe the marked region while suppressing any mention of the annotations. **Bottom Row (Cropping):** The input is a minimal crop tightly encompassing the annotated region. The LMM is given a general instruction (shown below the images) to describe the presented image, implicitly forcing focus onto the region of interest.

that Patch-ioner benefits substantially from the wider linguistic and visual variability captured in ReLaion, despite its noisier nature.

(2) Gains are not uniform across tasks. While the Trace and Dense tasks benefit the most from ReLaion (e.g., +2.4 CIDEr on Trace and +1.7 on Dense when moving from COCO to ReLaion 5.6M), the COCO-based Region-Set task exhibits a more nuanced behavior. The larger ReLaion sub-

Captioning Task: (Dataset)		Trace (COCO)		Dense (VG v1.2)		Region-Set (COCO Entities)		Image (COCO)		
Textual Decoder	# Parameters	C	P	C	P	C	P	C	P	CLIP-S
GPT2	124M	**27.9**	**78.7**	**31.9**	**78.8**	**109.1**	**87.5**	**69.2**	**87.4**	72.8
Gemma 3	270M	23.5	78.4	25.5	78.5	98.7	**87.5**	61.0	87.2	73.2
Qwen3	0.6B	23.5	78.4	25.0	78.4	98.7	**87.5**	59.7	87.2	73.4
LLama 3.2	1B	24.2	78.3	26.0	78.4	101.0	87.4	60.3	87.1	73.5
Qwen3	1.7B	22.9	78.2	24.5	78.2	95.6	87.4	57.9	87.1	**73.7**

Table 4. **Training different decoders.** CIDEr (C) and RefPAC-S (P) across four captioning tasks. The model adopted is T2D + Memory (≈ DeCap) trained on COCO train Karpathy split.

set (28.3M) significantly boosts Region-Set CIDEr, suggesting that recognizing localized entities and relations requires broader caption diversity, which becomes available only at larger scale.

(3) Memory Bank from ReLaion is less effective than the one from COCO. When employing 500k captions randomly sampled from ReLaion as memory bank, performance drops across many tasks. This suggests that using a collection of texts for the projection mechanism sampled from COCO leads the model to provide captions that are closer to the COCO ground-truth ones, particularly from a syntactic standpoint. In fact, we can notice how the larger performance drop is on the CIDEr metric.

Overall, these findings show that Patch-ioner adapts well to large-scale noisy text, outperforming COCO-trained baselines on most tasks, and that the advantages grow with the size of the ReLaion subset. This demonstrates that our framework can leverage broad web-scale text distributions to improve zero-shot captioning, even when training the textual decoder for a single epoch. However, we pick the model trained on COCO as reference to have a fair comparison with existing models.

11. Additional Results: Aggregation Strategies, Input Resolution, Text Collection

We tested several patch aggregation strategies and input resolutions for our model and the other baselines.

Patch Aggregation. In cases where we are not captioning a single patch, we test different aggregation functions for merging the $\mathbf{v}_i$ in a selected set S of visual patches:

a) **uniform**, the average box patch representations;
b) **gaussian**, for rectangular configurations of contiguous patches — i.e., either the full image or a bounding box; we consider a weighted average of patches representations where central patches weigh more; specifically, we assign to each patch (a, b) coordinates in a uniform square grid $[-1, 1]^2$ (i.e., the top-left and bottom-right patches have $(-1, -1)$ and $(1, 1)$ coordinates, respectively), and a weight of $e^{-(a^2+b^2)}$ in the average, and
c) **attention**, a weighted average of box patches representations, with patch weights defined as the average attention map of the last layer of ψ_{v}.

Input Resolution. For patch-based captioning, we followed Talk2DINO [5] and used an input image resolution of 518x518, obtaining 37 14x14 patches per side when using the Talk2DINO backbone. The original DeCap [33], ViECap [15], MeaCap [68] implementations uses the CLIP B/32 backbone with 224x224 input images with 7 patches per side. We also tested with the CLIP B/16 backbone, resulting in 14 patches per side at 224x224 resolution, and with 592x592 input image size, to obtain the same number of patches as in our framework (37 per side).

We report results of these additional configurations for all baselines: DeCap, ViECap, and MeaCap. While the main paper reports only the best configuration per task and model, in this section, we report and discuss the results of all the tested configurations. We perform these tests on COCO-derived datasets and on VG v1.2 for dense captioning. We highlight the rows in the tables corresponding to the configurations reported in the main paper.

Trace Captioning. Table 6 reports trace captioning results. We did not apply the *gaussian* weighting scheme for this task, as the sparse discontinuous traces often do not identify a rectangular region needed to apply this scheme. We notice that a) the simple average of the trace patches provides the best performance in our framework b) as expected, using the CLIP B16 backbone, that extracts finer patches, improves over the standard CLIP B32 backbone used in the baseline methods, and c) resolutions higher than 224 only marginally improve performance for baselines in this task.

Dense Captioning. Table 7 reports the results of the dense captioning task. For our framework, changing the weighting

Captioning Task: (Dataset)	Trace (COCO)		Dense (VG v1.2)		Region-Set (COCO Entities)		Image (COCO)		
Text-only Training Dataset	C	P	C	P	C	P	C	P	CLIP-S
COCO Train	27.9	78.7	31.9	78.8	109.1	87.5	69.2	87.4	72.8
ReLaion 5.6M	**30.3**	**79.0**	**34.1**	**79.0**	109.0	87.7	69.3	87.5	72.5
ReLaion 28.3M	29.7	**79.0**	33.6	78.9	**113.5**	**87.9**	**70.6**	**87.7**	**73.0**
Using a Memory Bank of 500k ReLaion Captions									
Relaion 5.6M	26.7	78.4	33.8	79.2	86.3	85.3	54.1	85.0	70.1
ReLaion 28.3M	26.8	78.3	33.4	79.0	86.6	85.3	54.6	85.0	70.5

Table 5. **Training on different datasets.** CIDEr (C) and RefPAC-S (P) across four captioning tasks. The model adopted is T2D + Memory ($\approx$ DeCap) using the GPT2 textual decoder.

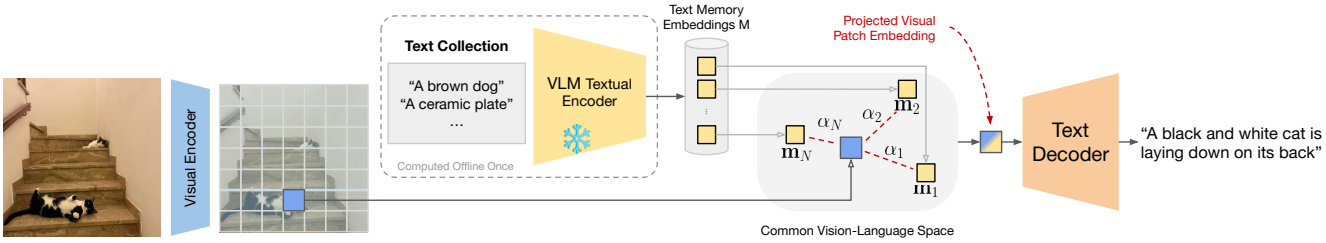

Figure 4. **Patch-level Captioning**. Given an input image, we first extract dense patch-level representations using a vision transformer backbone. For a selected patch, we apply the projection-based mechanism introduced by [33] to mitigate the modality gap and align its representation with the text embedding space. Finally, the transformed embedding is fed into a text decoder trained on a text-only corpus, generating a zero-shot caption for the patch.

strategy does not cause significant performance changes. The best baselines are the ones Region-based, which consist of applying the captioners to the CLS tokens of image crops specified by the bounding boxes (e.g., DeCap@224 Crop). The difference between the CLIP B/16 and B/32 versions is usually small or negligible.

Region-Set Captioning. Table 8 shows the results in the region-set captioning task on COCO Entities. The gap between zero-shot image captioners and our framework's captioners narrows in this task due to the more global nature of it, which requires the model to produce a caption for the whole image while focusing on certain regions. Also in this task, the choice of weighting strategy only marginally affects the performance of the models in our framework.

Image Captioning. In Table 9, we report the results of standard zero-shot image captioning. In addition to the already described weighting schemes, we test one additional configuration for our framework that is *central patch*, where the decoding is applied to the central patch of the image. We can observe that the most effective strategy for the image captioning task is *attention*. This is coherent with results from [5], where they suggest the attention-weighted patch means to use Talk2DINO for global tasks such as image-text retrieval.

Memory Bank. Considering that in the memory-based model of our framework (that is similar to [33]) we tackle the modality gap through a projection based on a collection of texts, we tested how much the selection of the texts in the memory bank influences performance. In Table 9, we also report the results obtained by that model when in its memory bank there are also ground-truth captions of the test set (rows marked with *GT Memory*). This provides a sort of upper bound to performance when varying the text collection used as memory. We observe that in this configuration, the performance only slightly improves (+0.5%), indicating that the model is robust to the choice of the memory bank.

12. Modality Gap: Projection to Textual Space vs Training with Noise

In this section, we quantitatively assess the performance of two state-of-the-art solutions to overcome the modality gap. In particular, we compared the configuration based on a memory bank of texts — the one introduced in §3 — with an alternative solution based on noise injection during the decoder training. Additionally, we include in our comparison a baseline with no modality gap mitigation (no mitig.), to highlight the benefits brought by each strategy.

Model	# Patches	Backbone	Input	Weighting	B	M	R	C	S	P
Image-based										
DeCap@224	7	CLIP B32	CLS	-	2.1	9.7	21.7	21.1	8.8	75.2
DeCap@224	14	CLIP B16	CLS	-	2.2	9.8	21.8	21.3	8.7	75.4
DeCap@592	37	CLIP B16	CLS	-	2.0	9.6	21.5	20.5	8.7	75.3
ViECap@224	7	CLIP B32	CLS	-	2.5	9.8	22.4	24.8	9.3	74.1
ViECap@224	14	CLIP B16	CLS	-	2.5	9.9	22.3	24.7	9.5	74.5
ViECap@592	37	CLIP B16	CLS	-	2.3	9.6	22.1	24.3	9.5	74.4
MeaCap@224	7	CLIP B32	CLS	-	2.3	9.4	21.5	23.1	8.9	74.2
MeaCap@224	14	CLIP B16	CLS	-	2.3	9.3	20.8	23.4	9.0	74.6
MeaCap@592	37	CLIP B16	CLS	-	2.2	9.1	20.3	22.5	9.0	74.4
Patch-based (Our Framework)										
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	uniform	**2.5**	**10.7**	**23.2**	**27.9**	**12.6**	**78.7**
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	attention	2.4	10.4	22.7	27.6	12.0	78.1

Table 6. Trace Captioning results on COCO test set.

Model	# Patches	Backbone	Input	Weighting	mAP	M	B	R	C	S	P
Image-based											
DeCap@224	7	CLIP B32	CLS	-	0.15	8.40	0.94	15.61	19.38	9.38	73.71
DeCap@224	14	CLIP B16	CLS	-	0.14	8.48	0.95	15.70	19.11	9.40	73.94
DeCap@592	37	CLIP B16	CLS	-	0.15	8.37	0.92	15.67	18.53	9.26	73.91
ViECap@224	7	CLIP B32	CLS	-	0.13	8.25	1.02	16.06	24.18	9.97	73.03
ViECap@224	14	CLIP B16	CLS	-	0.14	8.30	1.01	15.86	23.81	9.91	73.49
ViECap@592	37	CLIP B16	CLS	-	0.14	8.17	1.00	15.86	23.26	9.82	73.35
MeaCap@224	7	CLIP B32	CLS	-	0.13	8.04	0.98	15.40	23.22	9.66	72.97
MeaCap@224	14	CLIP B16	CLS	-	0.13	8.01	1.03	15.15	23.37	9.49	73.55
MeaCap@592	37	CLIP B16	CLS	-	0.13	7.86	0.99	15.13	22.77	9.43	73.50
Region-based											
DeCap@224 Crop	7	CLIP B32	CLS	-	0.17	10.03	1.35	18.20	23.61	10.90	77.09
DeCap@224 Crop	14	CLIP B16	CLS	-	0.18	10.33	1.40	18.44	24.56	11.28	77.76
DeCap@592 Crop	37	CLIP B16	CLS	-	0.14	8.30	1.05	16.39	17.20	7.78	75.47
ViECap@224 Crop	7	CLIP B32	CLS	-	0.15	9.32	1.42	17.79	26.40	10.07	74.34
ViECap@224 Crop	14	CLIP B16	CLS	-	0.16	9.59	1.46	18.03	27.13	10.43	75.62
ViECap@592 Crop	37	CLIP B16	CLS	-	0.12	7.83	1.14	16.02	20.20	7.44	73.26
MeaCap@224 Crop	7	CLIP B32	CLS	-	0.15	9.64	1.46	18.00	28.62	10.98	75.08
MeaCap@224 Crop	14	CLIP B16	CLS	-	0.16	10.03	1.57	18.45	30.53	11.51	76.35
MeaCap@592 Crop	37	CLIP B16	CLS	-	0.12	7.93	1.19	16.17	21.32	7.86	73.69
Patch-based (Our Framework)											
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	uniform	0.21	10.63	1.36	18.59	31.94	15.03	78.82
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	gaussian	**0.22**	**10.82**	**1.43**	**18.82**	**32.80**	**15.48**	**79.14**
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	attention	0.21	10.31	1.27	18.17	30.58	14.72	78.69

Table 7. Dense Captioning results on VG v1.2 test set.

Training with Noise. Various works [18, 43] proposed zero-shot image captioning solutions based on noise injection during the training of the text decoder. Through this strategy, the trained decoders are more effective in understanding semantic representations, even when those are not coming from the text modality. To implement this strategy in our framework, we trained the textual decoder on the same collection of captions as for the memory bank-based configuration. We adopted Talk2DINO [5] textual space for the decoder input space, which is aligned to DINOv2 [44] with registers [24]. Following the setting of [18], we added Gaussian noise with $\sigma^2 = 0.08$ to the textual embeddings while leaving the other parameters unchanged (as defined in §6). In the next paragraphs, we report and compare the results for each task of Talk2DINO within our framework with the memory bank ($\mathcal{M}emory$) and with the training with noise ($\mathcal{N}oise$).

In Table 10, we compare the two modality gap mitiga-

Model	# Patches	Backbone	Input	Weighting	B	M	R	C	S	P
Image-based										
DeCap@224	7	CLIP B32	CLS	-	10.1	19.0	38.0	94.4	26.4	86.9
DeCap@224	14	CLIP B16	CLS	-	10.0	19.4	38.3	95.1	26.8	87.4
DeCap@592	37	CLIP B16	CLS	-	9.6	18.6	37.5	91.4	25.9	86.7
ViECap@224	7	CLIP B32	CLS	-	11.2	18.2	38.9	102.7	27.0	85.0
ViECap@224	14	CLIP B16	CLS	-	11.3	18.3	38.6	102.2	26.9	85.4
ViECap@592	37	CLIP B16	CLS	-	10.8	17.8	37.9	99.2	26.5	85.0
MeaCap@224	7	CLIP B32	CLS	-	10.4	17.7	37.0	97.9	25.9	85.2
MeaCap@224	14	CLIP B16	CLS	-	10.1	17.5	35.5	96.5	25.7	85.4
MeaCap@592	37	CLIP B16	CLS	-	9.3	16.9	34.6	91.1	25.5	85.0
Patch-based (Our Framework)										
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	CLS	-	9.1	16.9	35.0	89.4	25.4	85.5
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	uniform	11.5	19.3	38.8	109.1	29.4	87.5
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	gaussian	**11.6**	**19.6**	**39.3**	**111.6**	**30.1**	**87.7**
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	attention	11.0	19.0	38.3	107.0	29.3	87.4

Table 8. Region-Set Captioning results for COCO Entities test set.

Model	# Patches	Backbone	Input	Weighting	B	M	R	C	S	P
Image-based										
DeCap@224	7	CLIP B32	CLS	-	23.46	25.12	50.06	87.40	19.14	90.58
DeCap@224	14	CLIP B16	CLS	-	**23.89**	**25.51**	**50.34**	**89.64**	**19.52**	**91.05**
DeCap@592	37	CLIP B16	CLS	-	22.43	24.64	49.25	84.57	18.66	90.36
ViECap @224	7	CLIP B32	CLS	-	26.70	23.99	50.85	89.67	17.54	88.45
ViECap@224	14	CLIP B16	CLS	-	26.3	24.0	50.3	89.5	17.6	88.8
ViECap @592	37	CLIP B16	CLS	-	25.60	23.38	49.52	86.84	17.08	88.33
MeaCap@224	7	CLIP B32	CLS	-	24.57	23.12	47.68	86.66	17.27	88.76
MeaCap@224	14	CLIP B16	CLS	-	23.6	22.7	45.5	85.1	17.3	89.0
MeaCap@592	37	CLIP B16	CLS	-	22.01	21.87	44.72	80.84	16.69	88.41
Patch-based (Our Framework)										
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	central patch	15.68	18.46	40.84	55.53	12.66	84.26
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	uniform	19.52	21.49	44.88	69.19	15.59	87.36
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	gaussian	21.17	22.62	46.62	76.79	16.73	88.36
T2D + Mem. ($\simeq$ DeCap) @518	37	DINOv2 B14	Patches	attention	23.64	23.93	48.54	88.46	18.21	90.21
T2D + Mem. ($\simeq$ DeCap) @518 GT Memory	37	DINOv2 B14	CLS	-	23.58	23.54	47.71	85.67	17.86	89.53
T2D + Mem. ($\simeq$ DeCap) @518 GT Memory	37	DINOv2 B14	Patches	attention	25.66	24.77	49.83	93.87	19.09	90.70

Table 9. Image Captioning results on COCO test set.

tion strategies across multiple captioning tasks, and also report the performance of a baseline without any mitigation (no mitig.). The baseline consistently underperforms compared to both the $\mathcal{M}emory$ and $\mathcal{N}oise$ configurations, indicating that, like other contrastively learned image-text encoders [36], Talk2DINO is also affected by the modality gap. These results highlight the importance of explicitly addressing this gap to achieve strong captioning performance. For Trace Captioning, the $\mathcal{M}emory$ method is slightly more effective in the semantic metric RefPAC-S, while the $\mathcal{N}oise$ variant achieves marginally better scores in CIDEr, ROUGE-L, METEOR, and BLEU@4, with a minimal gap between the two approaches. In Dense Captioning, the $\mathcal{M}emory$ model consistently outperforms the $\mathcal{N}oise$ model across all metrics. Similarly, for Region-Set Captioning, both methods achieve strong results, but the $\mathcal{M}emory$ method shows a clearer advantage, particularly in tasks closer to the patch level. Finally, in Image Captioning, the performance gap between the two architectures narrows, especially on the Flickr30k test split. In this scenario, the $\mathcal{M}emory$ method performs significantly better when applied to the CLS token, whereas patch aggregation produces comparable results. However, the metrics reveal conflicting trends across different datasets.

Chosen Strategy. Based on the observed results, we selected the projection-based approach ($\mathcal{M}emory$) as the primary strategy for overcoming the modality gap in our framework. While the noise injection method ($\mathcal{N}oise$) yielded competitive performance across multiple tasks, the $\mathcal{M}emory$

Table 10. **Mitigation of Modality Gap.** Comparison of Memory-based Projection ($\mathcal{M}emory$) vs Noise-trained Decoder ($\mathcal{N}oise$) across tasks.

	Trace Captioning (COCO)						Dense Captioning (VG v1.2)							Region-Set Captioning (COCO Entities)						Image Captioning (COCO)						
Mitigation	B	M	R	C	S	P	mAP	M	B	R	C	S	P	B	M	R	C	S	P	B	M	R	C	S	P	CLIP-S
no mitig.	1.2	9.1	18.3	14.7	8.5	75.1	0.18	9.7	0.7	15.9	17.8	10.2	75.2	5.0	15.0	29.4	59.4	21.1	82.2	9.9	17.7	36.8	43.7	12.3	82.2	69.6
$\mathcal{N}oise$	**3.0**	**11.5**	**24.7**	**29.3**	12.3	78.1	0.20	10.4	1.2	17.8	26.3	12.6	77.0	10.5	18.4	37.2	97.5	26.7	85.6	19.6	21.5	45.4	65.5	15.5	86.2	70.9
$\mathcal{M}emory$	2.5	10.7	23.2	27.9	**12.6**	**78.7**	0.21	10.6	1.4	18.6	31.9	15.0	78.8	11.5	19.3	38.8	109.1	29.4	87.5	19.5	21.5	44.9	69.2	15.6	87.4	72.8

method demonstrated superior performance in dense captioning and region-set captioning, as well as a clear advantage when applied to the CLS token in image captioning. Given these trends, and considering the stability of the projection-based approach across different evaluation settings, we adopted $\mathcal{M}emory$ as the default configuration for our framework.

13. Trace Captioning Benchmark Generation

We construct our Trace Captioning dataset from the Localized Narratives dataset [48]. This dataset consists of mouse traces and their corresponding speech transcriptions, where annotators describe objects in images while moving the mouse pointer over them.

The initial dataset samples include timestamped mouse traces and are composed of multiple sentences that thoroughly describe the trace, with the generated descriptions following the order of the mouse movement. However, our task does not require strict temporal coherence. Instead, we aim to generate a single, concise caption that describes the specific area covered by the localized trace, rather than a multi-sentence description.

To achieve this, we split the descriptions into individual sentences and align the traces accordingly. We then refine the traces by removing intermediate periods caused by transitions between sentences, which often occur when the annotator moves to a different region of the image. Specifically, we trim each trace by removing the first and last 15% of points, eliminating these transitional segments.

Furthermore, we refine the captions by prompting the Llama3 8B model to rephrase the sentences, removing vague or subjective phrases such as "there is," "we can see," or "on the left of the image," and replacing them with concise, objective descriptions that refer specifically to the region covered by the trace. This rephrasing is crucial to ensure that each caption adheres to the standard format of image-captioning datasets and focuses only on the precise part of the image that the trace corresponds to. The LLM also helps identify and remove irrelevant sentences (e.g., "the image is blurred," "the image is edited"), which are then discarded along with their associated traces from the final benchmark.

Figure 5 shows the full prompt used to guide the Llama model in refining and cleaning the descriptions. Figure 6 illustrates how the initial narrative samples are transformed into final trace captioning samples through the process of trace splitting and caption rephrasing.

14. Learned Patch Aggregation via Attention

In the main paper we employ a parameter-free aggregation strategy to combine patch embeddings belonging to a region. While this choice preserves the zero-shot nature of our framework, we additionally explored whether a lightweight learned aggregation module could further improve the quality of region representations when local supervision is available.

Attention-based aggregation. Let $S = \{v_i\}_{i=1}^{N}$, with $v_i \in \mathbb{R}^D$, denote the set of patch embeddings corresponding to a selected region. In the default formulation of our framework, the region representation is obtained through mean aggregation:

$$v_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} v_i. \tag{3}$$

To investigate a learned alternative, we introduce a lightweight attention-based aggregation module that summarizes the patch set through a single transformer-style attention layer. The patch embeddings are processed jointly and produce a summary token that attends over the region patches. Formally, given the matrix of patch embeddings $V \in \mathbb{R}^{N \times D}$, we compute

$$v_{\text{att}} = \text{Attn}(V), \tag{4}$$

where $\text{Attn}(\cdot)$ denotes a self-attention block that outputs a single `[CLS]` token representing the aggregated region information.

Rather than replacing the mean representation, we combine the two signals to preserve the stability of the parameter-free aggregation while allowing the model to learn refinements. The final region embedding is defined as

$$v_R = v_{\text{mean}} + \alpha\, v_{\text{att}}, \tag{5}$$

where α is a learnable scalar controlling the contribution of the attention-based summary. The parameter α is initialized to 0.1, ensuring that the aggregation initially behaves close to the mean operator and gradually learns to incorporate the attention-based refinement during training.

```
I have image descriptions derived from spoken narratives. These need to be rewritten as concise,
↪  stand-alone captions in the style of the image-caption datasets. Follow these rules:

- Remove unnecessary narrative phrases like "we can see," "there is," "in this image," etc.
- Ensure the caption is standalone and descriptive.
- Use simple, objective language that highlights key elements.
- Keep it concise--just a single phrase.
- Follow the classical style of caption datasets.
- If the description is vague, subjective, or does not describe a concrete visual element (e.g., "The
↪  image is taken indoor," "This image is blurred"), return `<INVALID>`.
- Wrap the output in `{}` and add nothing else.

### **Examples:**
- **Input:** "We can see a young elephant stands which is near the water in a wooded area."
  **Output:** {A young elephant stands near the water in a wooded area.}

- **Input:** "In this image I can see some young children kicking a soccer ball in a field."
  **Output:** {A group of young children kicking a soccer ball around a field.}

- **Input:** "In the left of the image, we see a pole that has two green street signs on it."
  **Output:** {A pole has two green street signs on it.}

- **Input:** "We can see two surfboards which are stuck in the sand along the seashore."
  **Output:** {Two surfboards stuck in the sand along the seashore.}

- **Input:** "This image consists of a man which rides a wakeboard behind a boat."
  **Output:** {A man rides a wakeboard behind a boat.}

- **Input:** "In the background, there are a bunch of sticky notes and a pair of scissors."
  **Output:** {A bunch of sticky notes and a pair of scissors.}

- **Input:** "It looks like a sepia-toned photograph of a motorcycle underneath the shadow of a
tree."
  **Output:** {A sepia-toned photograph of a motorcycle underneath the shadow of a tree.}

- **Input:** "There is a sky"
  **Output:** {A sky.}

- **Input:** "She is smiling."
  **Output:** {A smiling girl.}

- **Input:** "The image is taken indoor."
  **Output:** {<INVALID>}

- **Input:** "This image is edited."
  **Output:** {<INVALID>}

- **Input:** "The image is blurred."
  **Output:** {<INVALID>}

- **Input:** "I think he is about to jump."
  **Output:** {<INVALID>}

Now, rewrite the following captions accordingly. Wrap each in `{}` and add nothing else:
<INPUT CAPTION>
```

Figure 5. **LLM Prompt for rephrasing trace captions**.

Training setup. The attention-based aggregator is trained with region-level supervision for a single epoch using the COCO Trace Captioning training set derived from the LocNar COCO train split. We optimize the parameters of the attention layer and the scalar weight α using AdamW with learning rate 10^{-4}, while keeping the visual backbone and the text decoder frozen.

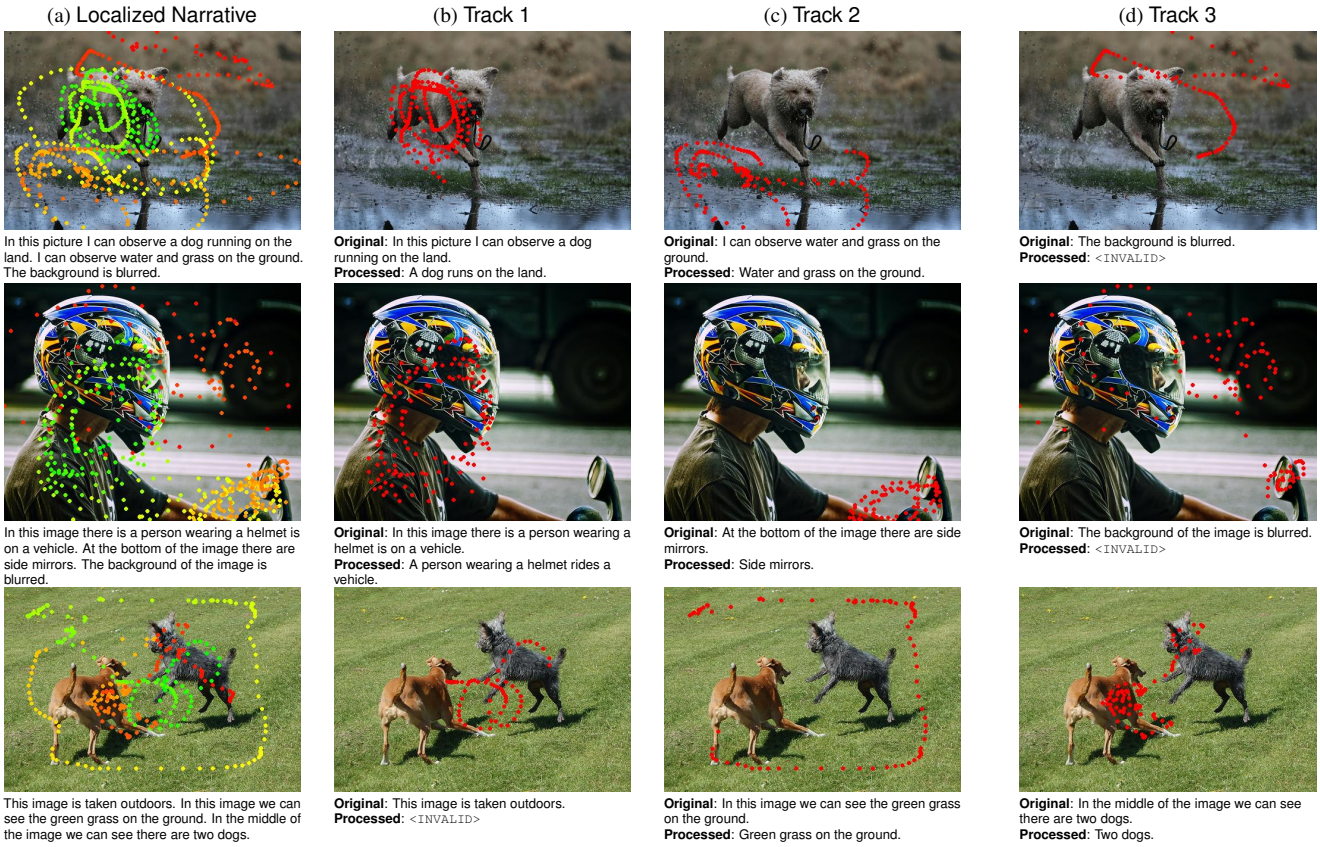

Figure 6. **Narrative vs. Trace Samples**. The first column displays sample images from the Localized Narrative dataset [49]. The remaining three columns show the corresponding mouse traces, along with the captions generated by the LLM. Captions marked with <INVALID> are removed from the dataset.

	Trace Captioning				Dense Captioning						Region-Set Captioning				Image Captioning					
	COCO		Flickr30k		VG v1.2			VG-COCO			COCO Entities		Flickr30k Entities		COCO			Flickr30k		
Model	C	P	C	P	mAP	C	P	mAP	C	P	C	P	C	P	C	P	CLIP-S	C	P	CLIP-S
Learned Aggregation (T2D + Noise)	**73.9**	**81.9**	**34.8**	**77.2**	17.8	**49.4**	**81.7**	17.7	**48.9**	**81.7**	67.8	83.6	25.7	73.3	**74.2**	86.5	70.9	27.3	79.7	66.0
Fixed Aggregation (T2D + Noise)	29.3	78.1	19.3	75.6	**20.3**	26.3	77.0	**20.3**	26.4	76.9	**97.5**	**85.6**	**37.1**	**76.5**	65.5	**86.2**	**70.9**	**27.8**	**80.8**	**67.0**

Table 11. Learned vs. Fixed Aggregation

Discussion. Table 11 reports the results of this experiment. The learned aggregator yields substantial improvements when evaluated on tasks with the same spatial granularity used during training (trace and dense captioning). However, these gains do not consistently transfer to tasks involving different region granularities (e.g., region-set or image captioning), where the fixed aggregation remains more balanced.

These findings suggest that when task-specific local supervision is available, learned aggregation can effectively refine region representations. However, in a unified zero-shot framework where no region-level annotations are assumed, the parameter-free aggregation strategy remains preferable due to its robustness across captioning granularities.

15. More Qualitative Results

Additional qualitative results are shown in Figures 7 and 8. Note that the first rows of Figures 7 and 8 contain also qualitative results for single patch captioning, for which we do not have annotated data to report quantitative results.

As can be noticed in Figures 7 and 8, the Region-Set Captioning task tends to align more closely with image-level captioning rather than strictly focusing on localized regions. This is expected since the ground-truth captions in the COCO Entities dataset originate from the image-level annotations of COCO, as stated in [10].

Figure 7. **Qualitative results.** We report four predictions of our model and compare baselines from the finer (top) to the coarser (bottom) task. For trace captioning examples, the trace time is color-coded from start (red) to end (yellow). **DeCap** = DeCap applied on the whole image. **DeCap (Crop)** = DeCap applied on cropped box. **ZeroCap** = ZeroCap [59] applied to the whole image. **CLOSE** = CLOSE [18] applied to the whole image. **Ours (CLIP + Mem.)** = Our patch-based framework using CLIP as backbone and the projection as modality gap mitigation strategy. **Ours (Talk2DINO + Mem.)** = Our patch-based framework using Talk2DINO as backbone and the projection as modality gap mitigation strategy. **GT** = ground-truth caption.

Figure 8. **Qualitative results.** We report four predictions of our model and compare baselines from the finer (top) to the coarser (bottom) task. For trace captioning examples, the trace time is color-coded from start (red) to end (yellow). **DeCap** = DeCap applied on the whole image. **DeCap (Crop)** = DeCap applied on cropped box. **ZeroCap** = ZeroCap [59] applied to the whole image. **CLOSE** = CLOSE [18] applied to the whole image. **Ours (CLIP + Mem.)** = Our patch-based framework using CLIP as backbone and the projection as modality gap mitigation strategy. **Ours (Talk2DINO + Mem.)** = Our patch-based framework using Talk2DINO as backbone and the projection as modality gap mitigation strategy. **GT** = ground-truth caption.