ProCap: Projection-Aware Captioning for Spatial Augmented Reality

Zimo Cao* Southwest University
Yuchen Deng† Southwest University
Haibin Ling‡ Westlake University
Bingyao Huang§ Southwest University

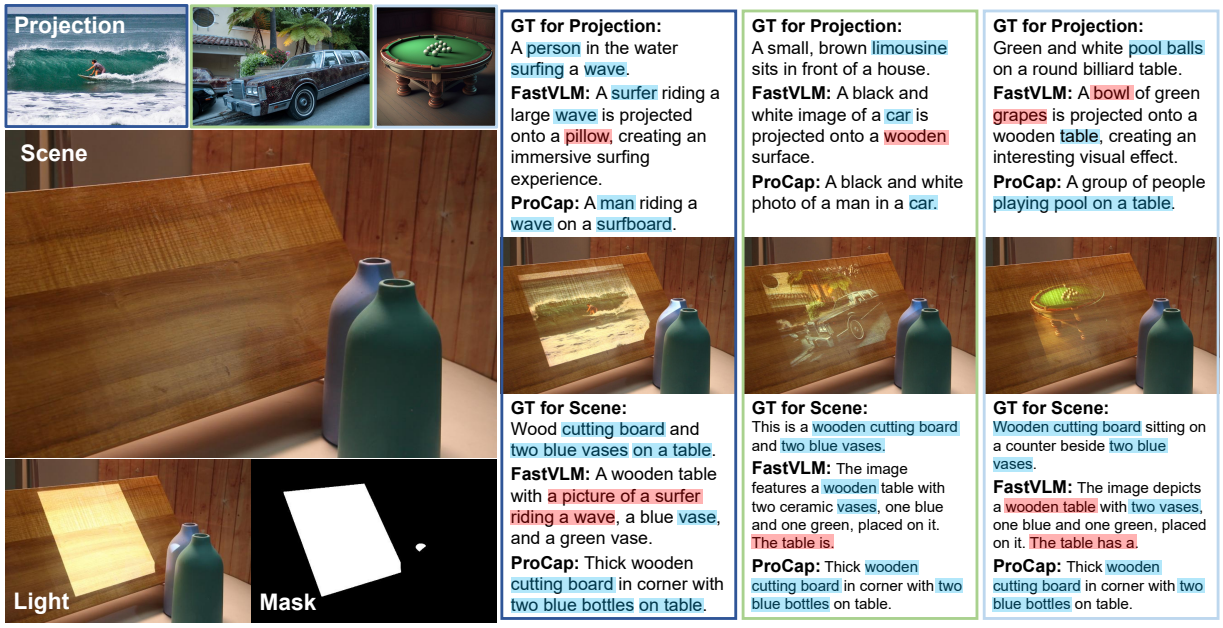

Figure 1: An overview of the **ProCap** framework for understanding complex SAR scenes with projected content. Given a challenging input image containing both **a physical scene** and **a digital projection** (left), our framework first employs an automatic segmentation module to disentangle the projection domain. Based on this mask, the model performs region-aware retrieval, sourcing distinct concepts for the physical environment and the projected content. Finally, our dual-task generation produces two independent and factually accurate descriptions: one for the **scene** and one for the **projection**. This approach successfully overcomes the critical failure of general-purpose VLMs, which often conflate the two contexts and generate a single, hallucinated caption. We highlight incorrect captioned objects in red and correct ones in blue. Note that the model outputs are truncated by the max tokens limit.

Abstract

Spatial augmented reality (SAR) directly projects digital content onto physical scenes using projectors, enabling immersive and collaborative experiences without requiring head-mounted displays. Yet, understanding such augmented scenes remains challenging for current vision language models (VLMs), which often lack spatial awareness, clear frames of reference, and dedicated datasets. To address these limitations, we introduce ProCap, a projection-aware captioning framework that integrates spatial and semantic knowledge for SAR scene understanding. At the core of ProCap is RGBP (RGB images with Projections), a large-scale dataset comprising 70 complex SAR scenes with diverse projections under varying lighting, occlusion, and object configurations. Building on this dataset, ProCap employs a two-stage framework: (1) automatic projection segmentation to separate projected virtual content from the physical environment, and (2) region-aware retrieval and attentive fusion to guide large language models in generating accurate and context-aware captions. Extensive experiments under a dual-task evaluation protocol demonstrate that our ProCap not only achieves competitive performance on physical scene description but also substantially outperforms state-of-the-art models on projection description. Ablation studies further validate the importance of segmentation and retrieval-based augmentation. Our work provides the first unified approach to SAR scene understanding, paving the way for multimodal interaction in projector-camera systems. The source code, pre-trained models and the RGBP dataset are available on the project page: https://ZimoCao.github.io/ProCap/.

Index Terms: Spatial augmented reality, vision language model, scene understanding.

1 Introduction

Spatial augmented reality (SAR) [8, 27] delivers digital content directly onto physical surfaces using devices like projectors rather than relying on screens. This technique has various applications in fields, including art design, industrial production, and user interaction. By using projected textures, colors, and other effects to improve the visual perception of objects, SAR enables multiple users to share and interact with the augmented scene simultaneously.

*e-mail: caozimo@email.swu.edu.cn
†e-mail: swudyc714@email.swu.edu.cn
‡e-mail: linghaibin@westlake.edu.cn
§e-mail: bhuang@swu.edu.cn. Corresponding author.

With the advancement of vision language models (VLMs) [49], a wide range of applications have emerged based on this technology, such as image-based question answering, visual grounding, and multimodal reasoning. However, applications specifically tailored to SAR remain largely unexplored. Several reasons contribute to this gap: (1) **Weak spatial awareness in SAR scene**. VLMs perform well in semantic recognition (identifying "apple", "chair", *etc*.) but often fail at spatial tasks like estimating depth, orientation, or surface curvature. This is critical in SAR, where projections must align with uneven surfaces or moving objects in real time, but VLMs are not trained for. (2) **Ambiguous frames of reference (FoR)**. VLMs frequently misinterpret spatial prepositions ("on", "under", "next to", *etc*.) unless given explicit guidelines. In projection systems, even if the surface area is correctly identified, misaligned FoRs lead to incorrect overlay placement. (3) Moreover, there are **few SAR datasets for VLMs**, leading to training data mismatch. Existing SAR evaluations focus on image quality assessment, projector-camera systems (ProCams) calibration, and *etc*. There is no unified metric to evaluate end-to-end SAR+VLM alignment quality, rendering reliability and user-perceived realism in one framework.

To address these issues, we propose ProCap, a novel framework designed to bridge this gap by explicitly disentangling the physical scene from the projected content. Our approach introduces a crucial first stage: an automatic segmentation module that identifies and isolates the projection area from the surrounding environment. This provides the VLM [39, 49] with unambiguous spatial context, directly tackling the challenges of weak spatial awareness and ambiguous frames of reference. Furthermore, to address the lack of training data, we introduce the RGBP dataset (RGB images with Projections), a new, large-scale dataset specifically created for SAR scenarios. Unlike conventional datasets, it provides rich annotations, including precise segmentation masks and, critically, separate ground truth (GT) captions for both the physical environment and the projected content. Finally, to enable fair and comprehensive evaluation, we propose a dual-task evaluation protocol. This protocol assesses the model's ability to generate accurate captions for the physical scene and the projection independently, offering a granular, end-to-end measure of its performance in SAR contexts. Through this combination of a specialized architecture, a tailored dataset, and a targeted evaluation methodology, ProCap significantly enhances the VLM's capacity for nuanced understanding and caption of complex SAR environments.

Our work focuses on improving the caption and recognition ability of large language models for complex SAR scenes including projection information. Our contributions can be summarized in three aspects:

- **ProCap Framework:** We propose a novel two-stage architecture that explicitly disentangles projected content from the physical scene. By utilizing an automated segmentation module and region-aware retrieval, our method provides the VLM with unambiguous spatial context, resolving the frame-of-reference ambiguity inherent in SAR scenes.

- **RGBP Dataset:** We introduce the first large-scale benchmark for SAR-based vision-language tasks, featuring 70 diverse physical environments paired with over 180,000 digital projections. The dataset provides dense annotations, including precise segmentation masks and decoupled ground-truth captions for both physical and projected layers.

- **Dual-Task Evaluation Protocol:** We establish a new evaluation paradigm that employs task-specific tokens to separate scene-level and projection-level descriptions. This protocol enables a granular, independent assessment of model performance, preventing the metric bias caused by task confusion in complex projection scenarios.

2 Related Work

2.1 Spatial augmented reality

Spatial augmented reality (SAR) aims to project designed patterns onto physical objects altering appearance. In this way, the observer does not need to wear any special equipment to perceive the augmented scene, ultimately blurring the boundary between digital and physical information. To construct immersive SAR, projector/sensor technologies and human perception and cognition are required.

2.1.1 Projection mapping

Projection mapping (PM) [33, 44–47, 59] projects virtual contents onto real object surfaces using projector-camera systems, and is widely utilized in creative arts, industrial design [12, 25], and entertainment. Interactive PM extends visual display to manipulation of human perception. User gestures, movements and expressions are tracked by sensors [7, 48] and cameras [4, 14, 28, 31], enabling interaction with the projected contents. Recently, Erel *et al*. [14] proposed an enhanced ProCams for real-time dynamic projection of 3D content onto user's hands. As diffusion models advance, instructional natural language is been fully utilized to control projected content [11, 13, 14]. Deng *et al*. [11] proposed language-guided projector image generation method with surface adaptation and stylization.

2.1.2 Plausibility illusion in SAR

Plausibility illusion (Psi) refers to the convincing sense that projected content in SAR plausibly interacts with the physical world, leading users to perceive it as real [53]. To improve Psi, one solution is projector compensation, which aims to produce the viewer's desired projection effects by modifying the projector input patterns to eliminate the distortions influenced by the environment [59] and projected surface or object attribute, further developing more applications as diminished reality. Early methods focused on geometry and color distortions, such as geometric correction combined with camera radiometric calibration [51]. While effective, these approaches often required re-calibration when projector settings changed, limiting flexibility under frequent adjustment. To address these issues, pixel-wise photometric compensation [16] using RGB image sampling and thin plate splines (TPS) interpolation over radiometric calibration. Under the limitation of the projector's physical brightness, computed compensation may exceed its capabilities, leading to artifacts. To achieve visually satisfactory projections, properties of the human visual system such as chromatic adaptation and perceptual anchoring are leveraged [26].

With the advent of deep learning, photometric [22] and geometric [21, 23, 24, 57, 58] compensation are formulated as an end-to-end learning problem. Leveraging optical illusions [43] to produce high-quality non-negative images explores compensation artifacts reduction. To better handle complex lighting scenarios such as multi-bounce illumination and to address the challenge of decomposing scene parameters, path tracing-based differentiable rendering [38] is utilized in ProCams, achieving superior performance in compensating dark textures. Furthermore, additional factors such as depth-induced blur and light occlusion still hinder the projection quality. Neural network-based methods remain effective on deblurring [32], and a lens-modified projection system [34] further removes the cast shadow in projection by condensing the projection light into a narrower beam. However, a disconnect remains between the plausibility illusion in SAR and VLM-based reasoning, causing a mismatch with human perceptual expectations. To our knowledge, no existing method addresses SAR scene comprehension through VLMs.

2.2 Domain-aware vision language model

Vision language models [2, 5, 6, 40, 49, 55, 64] aim to connect the two different modalities of vision and language and learn the correspon-

dence between them. Early VLMs achieve contrastive pre-training on large-scale image-text data [30,56], with strong cross-domain generalization, laying the foundation for subsequent generalized visual language tasks (*e.g.* image retrieval, text generation, *etc.*) [3, 5, 10]. Despite the excellent performance of generic models, directly using them for specialized domains suffers from limited effectiveness. To address this issue, solutions have been proposed for adaptation to specialized domains:

Full fine-tuning is commonly used for VLM domain adaptation. Researchers use small-scale labeled datasets from specific domains to further train the pre-trained general VLMs to adjust model parameters to better fit the tasks and data distribution of specialized domains. However, full fine-tuning of all model parameters is computationally expensive and often leads to overfitting, especially when only limited labeled data is available. To mitigate this, an intermediate step known as domain-specific continued pre-training is often introduced [18, 39]. In this approach, large-scale unlabeled or weakly labeled data from the target domain is used to continuously pre-train the model before downstream task fine-tuning. This allows the model to acquire domain-relevant visual features and linguistic patterns, thereby providing a more suitable initialization for subsequent fine-tuning and improving its effectiveness.

Parameter-Efficient Fine-tuning (PEFT) methods have been developed to further address the inefficiencies of full fine-tuning. Techniques such as adapter tuning [19] and low-rank adaptation (LoRA) [20] enable efficient adaptation by modifying only a small subset of parameters, significantly reducing computational cost while maintaining competitive performance. These methods add only a small number of trainable parameters or modules to the model while freezing most of the pre-trained model parameters. This not only greatly reduces the computational resources required for training, but also effectively avoids the problem of catastrophic forgetting on general knowledge.

Retrieval Augmentation (RA) addresses specialized domains with precise knowledge (*e.g.*, legal, medical). Some research efforts have attempted to combine language models with an external knowledge base [29, 36, 60]. When the model needs to answer a specialized question, it first retrieves relevant documents or cases from the knowledge base, and then uses the retrieved information as a context to generate an answer together with the input visual information.

In summary, although SAR research focuses on projection quality and plausibility illusion, and VLM studies explore domain adaptation and retrieval augmentation, none address this unique challenge of SAR scene understanding. The main obstacle is the lack of benchmarks, as existing datasets (e.g., COCO [42], nocaps [1], WHOOPS! [9]) do not capture projection-related phenomena.

3 RGBP Dataset

VLM performance depends on the diversity of the training data. While datasets like COCO [42] are standard for general recognition, they do not account for the interaction between digital projections and physical surfaces common in SAR. To address this, we developed the **RGBP** dataset to train and evaluate VLMs in complex projection environments.

3.1 Data acquisition

We used a projector-camera system (ProCams) in a hemispherical setup (Fig. 2). We varied two main factors to ensure the dataset covers a wide range of real-world conditions:

Lighting: We used ambient light and an RGB light panel (YONGNUO YN300Air II) to create scenarios ranging from bright, uniform illumination to dim, high-noise environments.

Geometry: We changed the projector's angle and placed physical objects stochastically to create diverse geometric distortions and surface occlusions.

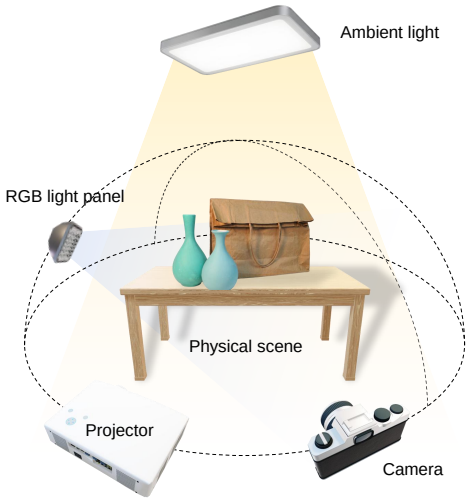

Figure 2: Configuration of the RGBP dataset capture environment.

Table 1: Detailed composition of the RGBP dataset. T_{proj} is the projection image set. Training data and 65 evaluation sets were captured using an EPSON ELPLP-78 (1600 × 1200) and a Canon 600D (1080 × 720). The remaining 5 evaluation sets used an EPSON EB-C2050WN (1600 × 1200) and a Nikon D3200 (1080 × 720).

Split	T_{proj} (per scene)	Scenes	Total Pairs
Training	COCO train $(2,000^*)$	60	**118,287** $(59 \times 2{,}000 + 287)$
Evaluation	COCO val (160) nocaps val (300) WHOOPS! (500)	70	**67,200** $(70 \times (160+300+500))$
Total		**70**	**185,487** (118,287 + 67,200)

*Scene60 only contains 287 images.

3.2 Composition of the training and evaluation dataset

The dataset composition are shown in Tab. 1, including hardware and resolution details. The projection image set, T_{proj}, pulls from COCO [42], nocaps [1], and WHOOPS! [9] for semantic variety.

Training set. We used 118,287 images from the COCO train split, distributed across 60 physical scenes. Each scene includes 2,000 unique projections (287 for the final scene, since scene60 only has 287 images). To help the model separate the physical scene from the projection, we also captured baseline frames for each scene using pure black, white, and gray projections. Then, we leveraged language models to assist in generating 50 high quality captions per scene. Five of these captions were randomly matched to each individual image. This strategy enhances both the diversity and accuracy of physical scene captions. Some training samples are shown in Fig. 3.

Evaluation set. The evaluation set includes 67,200 image pairs across 70 scenes, using 960 test images from COCO val, nocaps val, and WHOOPS!.

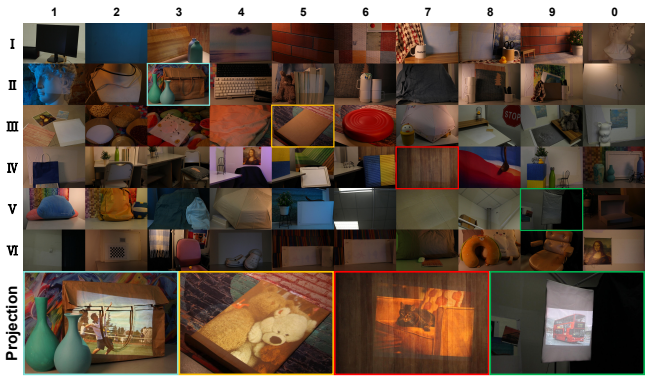

Figure 3: All 60 scenes used to train models. They are used to form the RGBP training dataset, represented by scenes 13 (blue box), 25 (yellow box), 37 (red box), and 49 (green box). The final data form of these four groups of representative scenes, combined with projected content, is also given below the 60 scenes.

3.3 Annotation and dual-task evaluation protocol

Instead of a single caption, RGBP uses a dual-task annotation scheme. For every image, we provide: (1) A binary segmentation mask of the projected area; (2) Two separate ground truth (GT) captions, i.e., a Scene GT (describing the room/objects) and a Projection GT (describing the projected image content).

A key innovation of the RGBP dataset is its specialized annotation scheme, which enables a granular, disentangled evaluation. For each generated image pair, we provide not only a precise binary segmentation mask but also two distinct GT captions, which form the basis of our dual-task evaluation protocol.

During evaluation, the model is prompted to generate two distinct captions. This allows us to measure how well the model understands the physical environment versus the projection overlay independently, providing a clearer picture of its performance than a single holistic score in SAR applications.

4 Projection-Aware Captioning (ProCap)

4.1 Problem formulation

The goal of ProCap is to distinguish the projected content from the physical scene despite distortions caused by color, brightness, and geometry. Existing VLMs face limitations in performing these two tasks simultaneously. We therefore formulate projection-aware captioning as a dual-task problem.

Given an image $\boldsymbol{I}$ captured from an SAR scene (Fig. 3), where $\boldsymbol{I}$ contains both physical objects and projected content, the model is required to generate two captions:

$$\left(C_{\text{phy}},\ C_{\text{proj}}\right) = \mathcal{F}(\boldsymbol{I}), \tag{1}$$

where C_{phy} denotes the caption describing the physical environment (e.g., objects, layout, lighting), and C_{proj} denotes the caption describing the projected content (e.g., text, images, animations).

The primary goal is to disentangle dual-source visual signals from $\boldsymbol{I}$ to generate contextually coherent, modality-specific descriptions. Formally, this interaction is modeled as a superimposition:

$$\boldsymbol{I} = \boldsymbol{I}_{\text{phy}} \oplus \boldsymbol{I}_{\text{proj}}, \tag{2}$$

where $\oplus$ represents the complex photometric and geometric blending of physical and projected content. Because direct disentanglement poses a significant challenge for standard VLMs, our ProCap framework approximates the inverse of this operation. By leveraging projection-aware segmentation and retrieval-augmented priors, we decompose the input space to facilitate the accurate generation of $(C_{\text{phy}}, C_{\text{proj}})$ within a unified captioning pipeline.

In order to complete the task of describing complex scenes containing projection information, we first need to clarify the difference between projection information and scene information. In ProCap framework, we roughly divide the model training into two steps: (1) separating the projected region and the physical scene region using segmentation, and (2) extracting features for the projected region and the remaining scene region, respectively.

4.2 Projection segmentation

The primary goal of this stage is to identify and segment projection areas within an input image $\boldsymbol{I}$. To formalize this process, we define the segmentation module as a function $\mathcal{S}$. This function represents a neural network trained specifically to differentiate projection content from its surrounding environment. The function takes an image $\boldsymbol{I}$ as input and produces a binary mask, M_{bin}, which mathematically represents the segmentation result.

$$M_{\text{bin}} = \mathcal{S}(\boldsymbol{I}) \tag{3}$$

In this output mask, pixel values of 1 represent the projection area, while values of 0 represent the physical scene area. This allows us to logically define the projection region (T_{proj}) and the scene region (R_{scene}) for the subsequent stage.

We adopt a coarse masking strategy to balance segmentation accuracy with computational robustness. Rather than computing instance-specific boundaries, we define R_{proj} using a standardized white-light projection for each scene. This approach intentionally discards peripheral edge noise and only keeps the high-entropy central region, effectively acting as a spatial regularizer. In this way, we prevent the model from overfitting to scene-specific artifacts and minimize the impact of environmental noise during the segmentation phase.

4.3 Region-aware retrieval and captioning

We use the projection obtained segmentation mask above to query an external knowledge base in a region-aware manner and generate the final description.

4.3.1 Unified external visual-name memory

Our framework utilizes a single, unified external knowledge base, structured as a key-value memory for efficient similarity-based retrieval. This knowledge base, denoted as $\mathcal{M}$, contains a large vocabulary of objects derived from datasets LVIS [17]. We formally define it as a set of n tuples, where each tuple consists of a visual embedding (key) and its corresponding name (value):

$$\mathcal{M} = \{(k_i, v_i)\}_{i=1}^{n} \tag{4}$$

Each key k_i is a high-dimensional vector representing the visual features of an object, and each value v_i is the textual name of that object.

4.3.2 Feature extraction and retrieval

First, a vision encoder $\mathcal{E}$ processes the image $\boldsymbol{I}$ to extract high-resolution feature maps, Z'_{hr}. To create distinct representations for the segmented regions, we then apply a Mask Pooling operation. This technique aggregates the features in Z'_{hr} only within the areas defined by our binary mask. For the projection region (M_{bin}), we compute a feature vector F_{proj} by averaging the corresponding features.

$$F_{\text{proj}} = \text{MaskPool}(Z'_{\text{hr}}, M_{\text{bin}}) \tag{5}$$

Similarly, we compute a distinct feature vector F_{scene} for the physical scene region by applying the same pooling operation to the areas where the mask value is 0.

$$F_{\text{scene}} = \text{MaskPool}(Z'_{\text{hr}}, \sim M_{\text{bin}}), \tag{6}$$

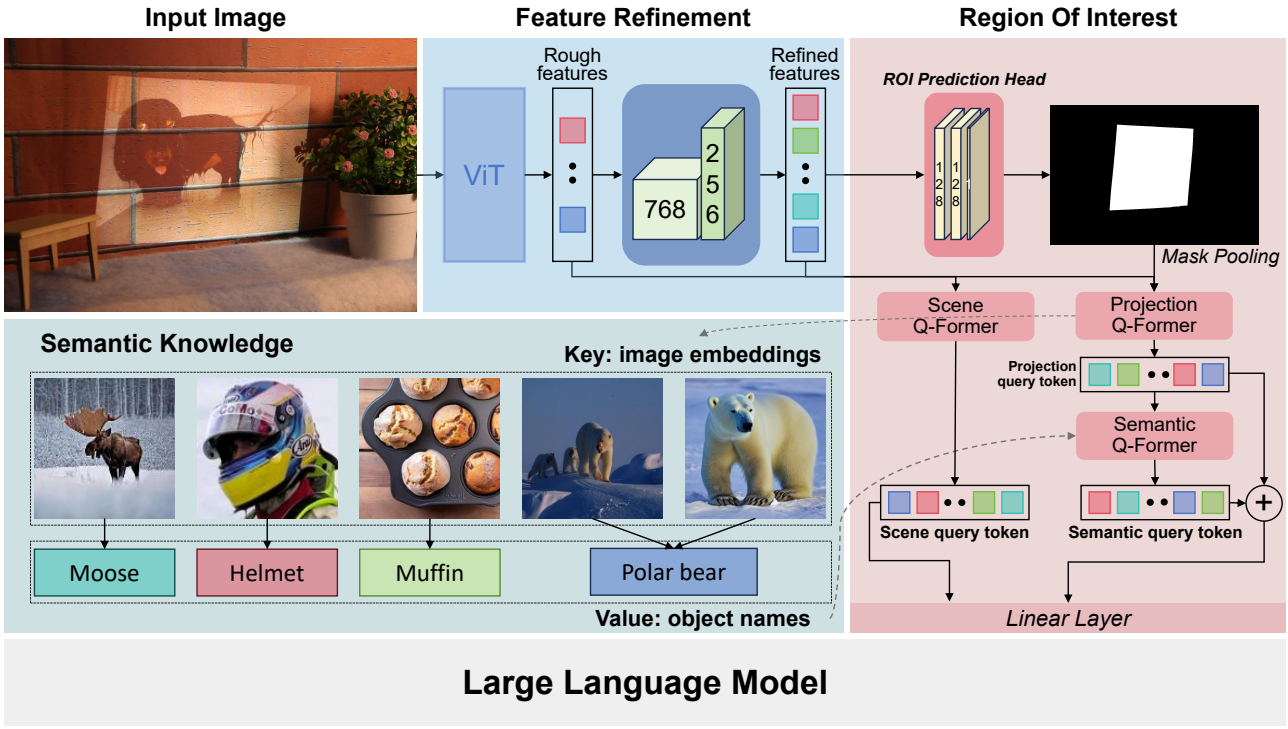

Figure 4: Overview of the proposed **ProCap** architecture. Given an input image containing both physical scene and projected content, a frozen vision transformer (ViT) backbone first extracts coarse visual features, which are refined into higher-resolution representations by a feature refinement module. An region of interest (ROI) prediction head estimates a projection mask, enabling **mask pooling** to disentangle scene and projection features. The disentangled features are then processed by two specialized Q-Formers: a **scene Q-Former** encodes global scene features, while a **projection Q-Former** focuses on projection-specific features. Retrieved semantic knowledge (object names) from an external memory is further encoded by a **semantic Q-Former** and fused with projection tokens. Finally, the scene and projection representations are projected into the embedding space of a frozen large language model via a linear connector, conditioned on task-specific tokens, to generate separate and accurate captions for both the physical scene and the projected content.

where operator $\sim$ is bitwise inversion. With these region-specific feature vectors, we can now query the unified knowledge base. We define a retrieval function $\mathcal{R}$, which takes a feature vector F, the knowledge base $\mathcal{M}$, and an integer K as input. The function works by calculating the cosine similarity between F and all keys k_i in $\mathcal{M}$, and returns the corresponding names v_i for the Top-K most similar keys. This process is applied in parallel to get two distinct lists of object names.

$$N_{\text{proj}} = \mathcal{R}(F_{\text{proj}}, \mathcal{M}, K) \tag{7}$$

The list N_{proj} contains the names of objects identified **within** the projected content. The Q-Former, denoted as $\mathcal{F}(\cdot)$, aggregates the feature vector F into a fixed number of global visual query features $\mathcal{Q}$. $\mathcal{Q}_0$ is a learnable initial query.

$$\mathcal{Q} = \mathcal{F}(F, \mathcal{Q}_0) \tag{8}$$

4.3.3 Multi-source fusion and generation

To generate a coherent description, the model must synthesize the global visual context with the specific information retrieved for each region. Therefore, an attentive fusion module, $\mathcal{F}_{\text{fusion}}$ is applied. This module takes two distinct inputs: the global visual query features $\mathcal{Q}$, the list of retrieved projection names N_{proj}. Through a series of cross-attention layers, it produces a single, contextually rich feature vector, H_{fused}.

$$H_{\text{fused}} = \mathcal{F}_{\text{fusion}}(\mathcal{Q}, N_{\text{proj}}) \tag{9}$$

The fused feature vector H_{fused} exists in a custom feature space. To make it compatible with the LLM, it must be projected into the LLM's word embedding space. We use a trainable linear layer, denoted by ϕ, for this transformation. The resulting vector, H_{prompt}, serves as the final, comprehensive prompt embedding for the decoder.

$$H_{\text{prompt}} = \phi(H_{\text{fused}}) \tag{10}$$

Finally, a frozen LLM decoder autoregressively generates the final caption $C = (c_1, ..., c_L)$ based on the prompt embedding H_{prompt}.

4.4 Training objective

The entire framework is trained end-to-end using a standard autoregressive language modeling objective. The goal is to maximize the likelihood of the model generating the GT caption $C = (c_1, ..., c_L)$, conditioned on the visual and retrieved information encapsulated in H_{prompt}. This is achieved by minimizing the negative log-likelihood (cross-entropy loss), $\mathcal{L}$, over the model's trainable parameters θ.

$$\mathcal{L}(\theta) = -\sum_{t=1}^{L} \log P(c_t | H_{\text{prompt}}, c_{1:t-1}; \theta) \tag{11}$$

The trainable parameters θ include the segmentation module $\mathcal{S}$, the global visual query features $\mathcal{Q}$, the fusion module $\mathcal{F}_{\text{fusion}}$, and the linear layer ϕ.

5 EXPERIMENTS

5.1 System configuration

We implemented and evaluated our ProCap based on different open-source large language models: TinyLlama-1.1B [61], OPT-2.7B [62], OpenLLaMA-3B [54], Vicuna-1.5-7B [63] for scalability. In addition, we fine-tune Qwen3-VL-8B-Instruct [5] on our RGBP dataset using supervised fine-tuning (SFT) and LoRA, following the procedure in [35]. Key training parameters were kept consistent across experiments: Training and inference were conducted on one NVIDIA A100, one NVIDIA RTX PRO 6000 Blackwell, and three NVIDIA RTX 4060Ti GPUs. More details are in the supplementary.

Compared methods. We compare ProCap against powerful baseline models from the FastVLM (0.5B, 1.5B, and 7B parameters) [55] and Qwen3-VL-Instruct (2B, 4B, and 8B parameters) [5]. They represent the state-of-the-art in efficient vision language modeling. For a fair comparison, the baselines were prompted with detailed instructions to perform the same dual-task description, that is similar to:

```
## For scene
Describe the scene in detail in the image, excluding any
projected content, as a short image caption.
## For projection
Describe any projected content in detail in the image,
excluding the surrounding scene, as a short image caption.
```

We also include our ProCap based on different parameters to analyze the impact of the base LLM's scale.

Evaluation metrics. We employ a standard suite of metrics to evaluate caption quality: BLEU@4, METEOR, CIDEr, and SPICE. As per our dual-task protocol, all scores are reported separately. For instance, we report CIDEr to divisively evaluate scene and projection captions to reflect the model's performance on each respective component of the scene. To independently evaluate scene and projection understanding, we introduce a dual-task evaluation protocol, which compares the model's generated descriptions for the physical scene and the projected content against their respective GT captions.

5.2 Results on domain benchmarks

We present a comprehensive comparison of ProCap against the baselines across our evaluation set. The results, organized by our dual-task protocol, clearly demonstrate the superiority of our approach, particularly in understanding projected content.

The performance on the 60 trained scenes, which includes COCO val subset, nocaps val subset, and WHOOPS! are shown in Tab. 2.

For the **scene captioning task**, our flagship ProCap $_{\text{TinyLlama-1.1B}}$ model achieves the highest scores across all metrics on COCO, with a CIDEr of 70.27, confirming its robust ability to understand physical environments. Furthermore, ProCap $_{\text{Vicuna-1.5-7B}}$ achieves a CIDEr of 36.39 on nocaps, while ProCap $_{\text{TinyLlama-1.1B}}$ once again achieves the top performance with a CIDEr of 69.95 on WHOOPS!, which is designed to test robustness against AI-generated projection content. This suggests the capability of our model to understand physical space is more robust and less susceptible to distraction from projection content. Further qualitative results are shown in Fig. 5. In contrast, the baseline models struggle significantly, indicating a difficulty in separating physical scene information from the complex SAR scene.

For the **projection captioning task**, result on trained scenes are presented in Tab. 2. Our ProCap establishes an even more significant advantage. While ProCap $_{\text{Vicuna-1.5-7B}}$ achieves a CIDEr of 78.99 on COCO, it also performs excellent on nocaps, which is an order of magnitude higher than the all the baselines. Moreover, even though all models find it challenging to describe bizarre content on WHOOPS!, our ProCap still provides a more coherent interpretation than the baselines, as evidenced by its higher scores. This proves the robustness and necessity of our specialized approach even in adversarial scenarios.

In the meanwhile, according to Tab. 3, on the first 5 untrained scenes captured with identical ProCams configurations, ProCap demonstrated only marginal superiority in the scene captioning task based on CIDEr metrics, failing to exhibit significant advantages. This limitation stems from model overfitting due to the limited training dataset of 60 scenes. However, after fine-tuning the Qwen3-VL-8B-Instruct model using the RGBP dataset, we observed that all metrics of RGBP $_{\text{Qwen3-VL-8B-Instruct}}$ significantly outperform the baseline models in projection captioning task. This demonstrates the significant advantage of RGBP dataset in prompting the projection captioning task in complex SAR scenes.

Furthermore, we evaluated the last 5 evaluation sets of the RGBP dataset using RGBP $_{\text{Qwen3-VL-8B-Instruct}}$, with both the vision encoder and language model adapted via LoRA. The last 5 evaluation sets were captured using ProCams configurations that differ from those in the training and the first 5 evaluation sets, as detailed in Tab. 1. As shown in Tab. 4, the performance on the new ProCams setups is nearly identical to that achieved on the original evaluation sets in Tab. 3.

5.3 Effectiveness of the semantic knowledge

The effectiveness of retrieval-augmented mechanisms in image captioning has been substantiated by EVCap [41], which demonstrates that an external visual-name memory significantly bolsters the capacity of LLMs to comprehend open-world objects. We further extend this verification to SAR scenes, which exhibit considerably higher complexity than traditional image captioning tasks. To isolate the impact of semantic knowledge, we compared the performance of our ProCap $_{\text{Vicuna-1.5-7B}}$ w/ or w/o augmentation, which replaces all semantic values with null. This setup ensures that while the retrieval mechanism based on visual features remains active, the model receives no semantic knowledge to assist in captioning.

As detailed in Tab. 5, the inclusion of semantic knowledge leads to substantial performance gains across all evaluated benchmarks. On the COCO set, the CIDEr score improves significantly from 67.98 to 86.26. A similar trend is observed in the nocaps set, where the overall CIDEr and SPICE scores rise by 7.66 and 1.62, respectively. Furthermore, on the WHOOPS! set, which focuses on commonsense-violating synthetic images, the performance improves from 20.48 to 24.44 in CIDEr. These results underscore that retrieval-based augmentation still remain a critical guiding factor for achieving accurate and contextually relevant descriptions in high-complexity SAR scenes.

5.4 Effectiveness of the segmentation module

To investigate the contribution of the segmentation-based architecture to ProCap, we first analyze the refinement module. As shown in Tab. 7, removing this module (*w/o refinement*) leads to a noticeable performance degradation, particularly in the Projection task where the CIDEr score on COCO drops from 66.03 to 52.55. This decline suggests that the refinement module plays a crucial role in extracting fine-grained visual features from segmented regions. While SPICE scores remain relatively stable in some cases-likely because the model still recognizes basic semantic categories-the significant drop in CIDEr indicates that the quality of sentence structure and precise visual-textual alignment heavily depends on these refined features.The role of explicit masking is further evaluated by comparing ProCap with a variant trained without masks (*w/o mask*). As reported in Tab. 6, the *w/o mask* version suffers a severe performance collapse in the Scene task for trained scenes (CIDEr 70.27 $\rightarrow$ 22.81). This demonstrates that without explicit spatial guidance, the model fails to distinguish between the background scene and the projection content, leading to severe feature interference. Interestingly, in untrained scenes, the *w/o mask* variant occasionally achieves higher SPICE or CIDEr in the Projection task. We attribute this to the fact that, without mask constraints, the model may lean towards a

Table 2: Performance comparison on **the 60 trained scenes**. Metrics are reported as BLEU@4 (B@4), METEOR (M), CIDEr (C), SPICE (S). Results are averaged on the 60 scenes, and the best results are in **bold**.

Task	Method	COCO Test				nocaps val In-domain		nocaps val Near-domain		nocaps val Out-domain		nocaps val Overall		WHOOPS! Test	
		B@4 ↑	M ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene caption	FastVLM-0.5B [55]	4.76	14.25	1.72	9.20	1.87	8.83	1.78	8.93	2.16	10.65	1.60	9.47	1.59	10.17
	FastVLM-1.5B [55]	4.75	14.15	1.92	8.77	1.98	8.44	1.96	8.64	2.33	10.42	1.73	9.16	1.74	9.92
	FastVLM-7B [55]	5.62	15.63	2.31	10.31	2.51	10.13	2.55	10.16	2.87	12.08	2.21	10.79	2.21	11.98
	Qwen3-VL-2B-Instruct [5]	8.58	19.18	2.34	12.17	2.42	11.79	2.48	11.82	2.72	12.89	2.12	12.17	2.01	12.83
	Qwen3-VL-4B-Instruct [5]	5.99	18.32	2.38	11.68	2.66	11.80	2.64	11.95	2.64	12.07	2.19	11.94	2.02	12.15
	Qwen3-VL-8B-Instruct [5]	7.19	19.75	2.38	13.03	2.50	13.20	2.50	12.99	2.60	13.24	2.09	13.15	1.97	13.07
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	**36.50**	**29.23**	**70.27**	21.92	**51.50**	21.54	53.99	23.38	**50.03**	23.21	35.21	24.21	**69.95**	21.87
	ProCap $_{\text{OPT-2.7B}}$ [62]	35.95	28.75	28.01	**24.34**	28.44	**24.44**	28.34	**24.46**	28.46	**24.48**	27.63	**24.45**	28.19	**24.11**
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	35.96	28.89	69.43	22.01	49.69	20.80	**55.63**	22.64	49.30	23.17	37.14	22.75	69.46	22.08
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	34.15	29.13	36.17	22.75	36.98	22.94	36.22	22.84	36.07	22.85	36.39	22.88	36.53	22.70
	RGBP $_{\text{Qwen3-VL-8B-Instruct}}$ [5]	36.38	29.18	37.81	23.22	37.41	23.47	37.80	23.48	37.32	23.54	**37.48**	23.49	37.70	23.59
Projection caption	FastVLM-0.5B [55]	4.58	14.18	7.17	7.43	6.91	6.39	6.93	6.32	3.89	4.35	6.16	5.68	5.54	5.85
	FastVLM-1.5B [55]	4.75	14.02	7.21	7.31	6.82	6.79	6.20	6.36	3.64	4.48	5.72	5.88	4.14	6.09
	FastVLM-7B [55]	5.26	14.15	7.65	7.33	7.11	6.77	7.30	6.88	5.91	4.69	7.01	6.12	6.25	6.77
	Qwen3-VL-2B-Instruct [5]	7.97	18.80	10.73	12.47	14.41	11.11	16.47	10.93	13.39	9.12	14.85	10.38	9.82	11.05
	Qwen3-VL-4B-Instruct [5]	5.85	18.18	10.59	11.72	14.64	11.14	17.06	10.90	13.25	9.37	14.93	10.47	10.35	10.91
	Qwen3-VL-8B-Instruct [5]	6.10	17.83	11.56	11.63	15.38	10.31	17.64	10.66	13.92	9.07	15.57	10.01	11.19	10.97
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	15.68	16.61	54.37	9.75	39.62	5.74	29.77	4.79	17.69	3.64	30.45	4.80	11.88	3.62
	ProCap $_{\text{OPT-2.7B}}$ [62]	18.05	16.83	57.72	10.21	46.48	6.91	30.40	5.41	13.03	3.19	30.58	5.17	15.04	4.40
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	24.18	20.59	76.98	13.58	57.61	8.43	42.73	6.82	24.10	4.57	42.06	6.66	19.37	5.43
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	24.58	21.05	78.99	13.83	55.19	8.32	39.75	6.68	22.85	4.75	39.93	6.59	22.33	6.02
	RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	**36.37**	**27.75**	**127.58**	**21.14**	**99.52**	**13.17**	**107.20**	**13.74**	**98.00**	**12.68**	**102.67**	**13.19**	**80.46**	**16.07**

Table 3: Performance comparison on **the first 5 untrained scenes**. Metrics are reported as CIDEr (C), SPICE (S). Results are averaged on the 5 scenes. The best results are highlighted in **bold**, and the second best results are underlined.

Task	Method	COCO Test		nocaps val Overall		WHOOPS! Test	
		C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene caption	FastVLM-0.5B [55]	1.44	5.42	1.32	5.58	1.02	6.78
	FastVLM-1.5B [55]	0.98	5.36	1.04	6.44	0.84	8.50
	FastVLM-7B [55]	1.24	6.00	1.38	6.72	1.34	8.66
	Qwen3-VL-2B-Instruct [5]	2.36	9.32	2.18	8.86	1.86	8.94
	Qwen3-VL-4B-Instruct [5]	3.44	9.08	3.42	8.88	2.98	9.04
	Qwen3-VL-8B-Instruct [5]	3.68	**12.26**	3.52	**12.00**	3.08	**12.46**
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	5.02	5.53	4.06	8.34	**7.48**	5.15
	ProCap $_{\text{OPT-2.7B}}$ [62]	**5.80**	9.44	4.82	9.12	4.68	8.90
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	3.78	7.77	4.02	7.10	5.73	6.60
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	2.94	7.50	2.34	7.30	2.22	7.12
	RGBP $_{\text{Qwen3-VL-8B-Instruct}}$ [5]	4.82	9.40	**4.96**	8.86	4.00	9.42
Projection caption	FastVLM-0.5B [55]	11.54	9.38	8.94	7.02	8.72	8.06
	FastVLM-1.5B [55]	10.00	8.92	6.84	7.26	4.88	7.70
	FastVLM-7B [55]	9.70	9.38	9.44	8.14	7.84	9.08
	Qwen3-VL-2B-Instruct [5]	14.10	16.14	19.60	12.84	13.04	13.78
	Qwen3-VL-4B-Instruct [5]	12.86	14.32	18.46	12.58	12.72	12.88
	Qwen3-VL-8B-Instruct [5]	14.34	14.68	19.20	12.34	13.90	13.10
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	46.90	12.60	32.36	5.10	15.23	4.57
	ProCap $_{\text{OPT-2.7B}}$ [62]	63.40	10.96	34.18	5.68	16.72	4.82
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	66.03	14.90	44.72	7.06	24.65	6.80
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	86.26	14.88	43.94	7.10	24.44	6.60
	RGBP $_{\text{Qwen3-VL-8B-Instruct}}$ [5]	**136.60**	**22.24**	**108.66**	**13.76**	**85.82**	**17.02**

Table 4: Performance comparison on **the last 5 untrained scenes**. Note that all the methods are based on Qwen3-VL-8B-Instruct. Metrics are reported as CIDEr (C), SPICE (S). Results are averaged on the 5 scenes, and the best results are in **bold**. "Proj." denotes Projection.

Task	Method	COCO Test		nocaps val Overall		WHOOPS! Test	
		C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene	w/ RGBP fine-tuned	**3.18**	6.84	**2.96**	6.80	**2.92**	7.46
	w/o RGBP fine-tuned	0.48	**9.84**	0.38	**9.58**	0.32	**10.20**
Proj.	w/ RGBP fine-tuned	**133.50**	**22.18**	**106.38**	**13.64**	**87.10**	**17.34**
	w/o RGBP fine-tuned	4.96	15.32	8.08	13.54	5.80	13.18

Table 5: Performance comparison on **the first 5 untrained scenes** w/ and w/o retrieval-based augmentation in projection captioning. Note that all the methods are based on ProCap $_{\text{Vicuna-1.5-7B}}$ [63]. Metrics are reported as CIDEr (C), SPICE (S). Results are averaged on the 5 scenes, and the best results are in **bold**.

Method	COCO Test		nocaps val Overall		WHOOPS! Test	
	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
w/ augmentation	**86.26**	**14.88**	**43.94**	**7.10**	**24.44**	**6.60**
w/o augmentation	67.98	11.64	36.28	5.48	20.48	5.50

"global-view" bias which happens to capture broader context; however, this comes at the cost of losing the ability to precisely describe the scene, proving that explicit segmentation is indispensable for a reliable dual-task system.Finally, we analyze the necessity of our dual-task design by comparing ProCap with single-task "specialists." As shown in Tab. 7, although a model trained exclusively on projections (*w/o scene Q-Former*) may show a marginal advantage in certain projection metrics (e.g., 67.38 vs. 66.03 CIDEr on COCO), it completely loses the capability to describe the surrounding scene. ProCap achieves a Pareto-optimal balance between the two tasks. Our segmentation-based disentanglement allows a single model to match the performance of specialized experts in both domains simultaneously. This synergy proves that ProCap does not merely "average" the two tasks but effectively decouples them at the feature

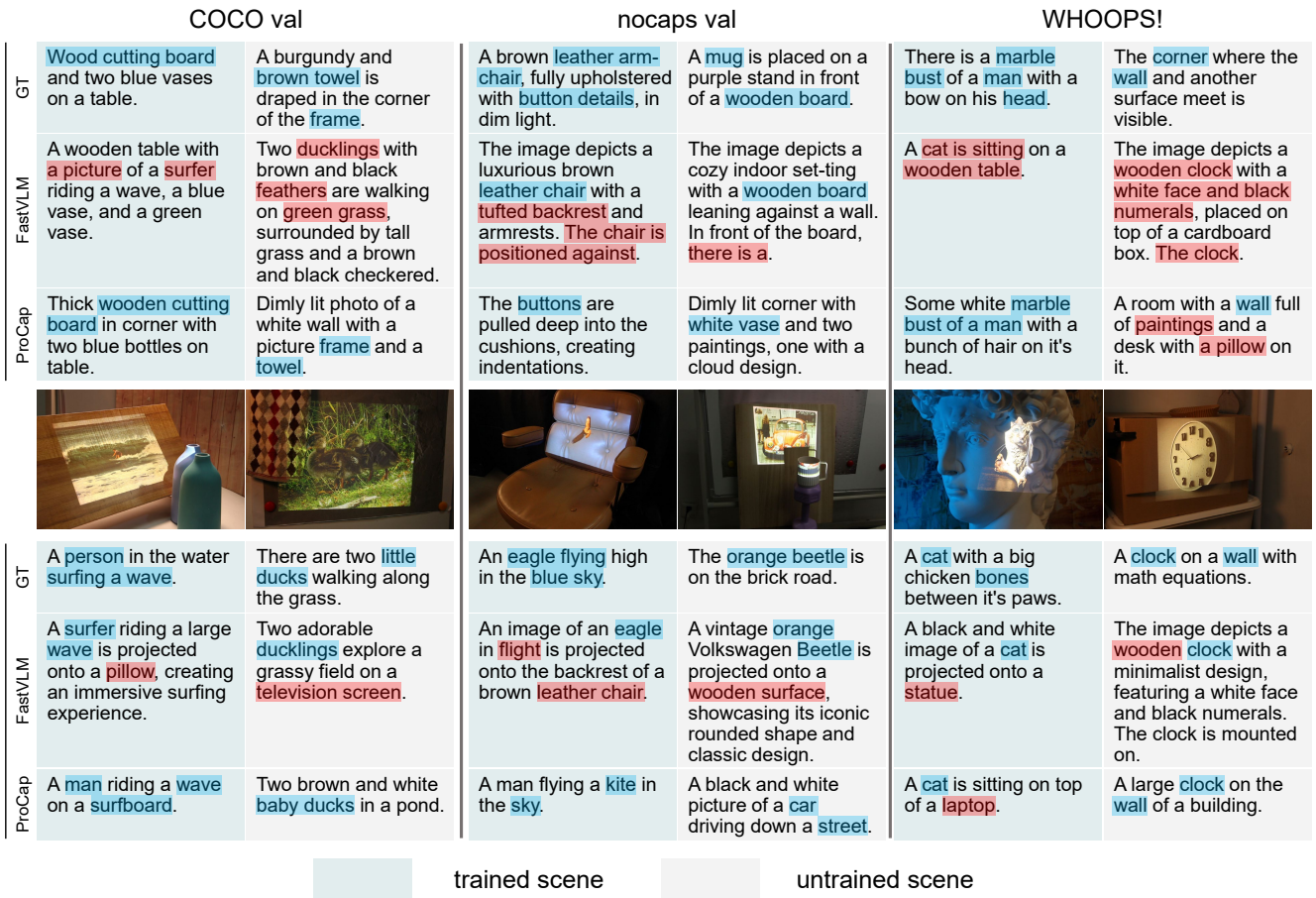

Figure 5: We compare the descriptive ability of ProCap $_{\text{OpenLLaMA-3B}}$ [15] with FastVLM-7B [55] in complex SAR scenes. First, we list all the corresponding data sets used for testing, they are **COCO val** [42], **nocaps val** [1], and **WHOOPS!** [9]. The performance of ProCap and the other two methods in 60 sets of scenes that have been trained and 5 sets of scenes that have not been trained are tested, respectively. Above the test picture is the comparison of the description ability of projection information, and the bottom of the test picture is the comparison of the descriptive ability of scene information. We highlight incorrect captioned objects in red and correct ones in blue. Note that the model outputs are truncated by the max tokens limit.

level, ensuring high-quality descriptions for both the projection and its environment without requiring task-specific backbones.

Table 6: Performance comparison on **the 60 trained scenes and the first 5 untrained scenes** w/ and w/o mask. Metrics are reported as CIDEr (C), SPICE (S). Results are averaged, and the best results are in **bold**. "Proj." denotes Projection.

Task		Method	COCO Test		nocaps val Overall		WHOOPS! Test	
			C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Trained scenes	Scene	ProCap $_{\text{TinyLlama-1.1B w/ mask}}$	**70.27**	21.92	**35.21**	**24.21**	**69.95**	21.87
		ProCap $_{\text{TinyLlama-1.1B w/o mask}}$	22.81	**23.10**	22.44	23.15	60.49	**23.05**
	Proj.	ProCap $_{\text{TinyLlama-1.1B w/ mask}}$	**54.37**	9.75	**30.45**	4.80	**11.88**	**3.62**
		ProCap $_{\text{TinyLlama-1.1B w/o mask}}$	52.79	**10.31**	24.47	**5.08**	10.93	3.61
Untrained scenes	Scene	ProCap $_{\text{TinyLlama-1.1B w/ mask}}$	**5.02**	5.53	**4.06**	**8.34**	**7.48**	5.15
		ProCap $_{\text{TinyLlama-1.1B w/o mask}}$	2.32	**6.36**	1.78	6.50	6.85	**7.00**
	Proj.	ProCap $_{\text{TinyLlama-1.1B w/ mask}}$	46.90	**12.60**	**32.36**	5.10	15.23	4.57
		ProCap $_{\text{TinyLlama-1.1B w/o mask}}$	**56.50**	10.68	25.58	**5.20**	**19.70**	**5.98**

6 Discussion

Results demonstrate effectiveness of ProCap and the RGBP dataset. This success stems from explicitly disentangling the physical scene and projected content, evidenced by the order-of-magnitude gains in projection captioning on COCO and nocaps. By segmenting the projection area first, our model describes content without interference from the surrounding environment. In contrast, holistic baselines often merge these features, leading to the hallucinations and factual errors noted in our qualitative analysis.

ProCap also shows robustness on the adversarial WHOOPS! dataset [9]. The model maintains stable scene grounding by treating illogical projections as a distinct layer of information rather than a source of confusion. Such resilience to distracting or nonsensical digital overlays is essential for reliable applications in AR and human-robot interaction.

However, our approach has limitations. The model's performance is contingent upon the accuracy of the initial segmentation module. In cases where projections are highly transparent, irregularly shaped, or subtly integrated into the environment, error propagation from segmentation failures could degrade the final description quality. Additionally, while our retrieval mechanism enhances descriptive detail, it is constrained by the vocabulary and concepts present in the LVIS-

Table 7: Performance comparison showing the effectiveness of different modules on **the first 5 untrained scenes**. Metrics are reported as BLEU@4 (B@4), METEOR (M), CIDEr (C), SPICE (S). Results are averaged on the 5 scenes, and the best results are in **bold**.

Task	Method	COCO Test				nocaps val In-domain		Near-domain		Out-domain		Overall		WHOOPS! Test	
		B@4 ↑	M ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene	ProCap $_{\text{OpenLLaMA-3B}}$	1.50	**11.73**	**3.78**	7.77	**4.88**	**7.16**	**4.96**	**7.28**	**5.56**	**7.14**	**4.02**	**7.10**	**5.73**	**6.60**
	ProCap $_{\text{OpenLLaMA-3B, w/o refinement}}$	0.25	10.60	3.68	**7.78**	4.56	6.10	4.34	6.28	4.62	6.46	3.60	6.28	5.52	5.73
	ProCap $_{\text{OpenLLaMA-3B, w/o projection Q-Former}}$	**1.88**	10.60	3.80	6.67	4.74	5.88	4.22	5.26	4.12	5.94	3.42	5.60	6.18	5.42
Projection	ProCap $_{\text{OpenLLaMA-3B}}$	19.90	21.22	66.03	14.90	58.22	8.66	**48.00**	7.34	**27.10**	5.20	**44.72**	7.06	**24.65**	6.80
	ProCap $_{\text{OpenLLaMA-3B, w/o refinement}}$	9.90	19.90	52.55	15.60	47.02	**9.00**	38.30	**7.44**	24.42	**5.38**	37.92	**7.26**	15.22	**7.75**
	ProCap $_{\text{OpenLLaMA-3B, w/o scene Q-Former}}$	**27.02**	**23.72**	**67.38**	**16.63**	**60.20**	8.70	46.66	7.22	23.54	4.24	44.22	6.76	22.18	6.57

based knowledge base proposed by EVCap [17, 41]. Describing highly specialized or novel content not covered by this knowledge base remains a challenge. Our experiments show that projection captioning exhibits strong adaptability from train to untrain scenes, yet a noticeable performance gap remains in scene captioning. This is likely due to the inherent challenges of SAR scenes, such as complex lighting and material properties, and the limited scale of real-world training data, despite capturing 60 scenes including 50 captions per scene, this is still small compared to over 118, 000 projection images and multiple captions per image in COCO [42]. Addressing these limitations could involve domain adaptation, few-shot learning, architectural improvements, or leveraging synthetic data to further enhance generalization and reduce labor cost.

6.1 Applications

The contributions of this paper extend beyond the immediate task of image captioning, positioning our work as a key enabler for future advancements in large-scale AI systems and providing a foundational resource for the research community.

6.1.1 ProCap as specialized expert in MoE models

The development of massive AI models is increasingly trending towards the Mixture-of-Experts (MoE) [37] architecture. In this paradigm, a monolithic model is replaced by a collection of smaller, specialized "expert" networks, with a lightweight gating network, or "MoE Router", dynamically routing tasks to the most appropriate expert. This allows for massive scalability while maintaining computational efficiency.

Our work, ProCap, is architecturally suited to function as such a specialized expert for the domain of Spatial Augmented Reality. Whereas general-purpose VLMs act as "generalists", ProCap provides a robust and accurate solution to the specific challenge of disentangling physical and digital content. We envision ProCap being integrated as an expert module within a future, large-scale multimodal MoE system. When a high-level MoE Router identifies an input image containing SAR elements, it could bypass the generalist VLMs and route the task directly to the ProCap expert. This would ensure that mixed-reality scenes are processed with the necessary precision, dramatically improving the overall model's robustness and accuracy in this increasingly important domain.

6.1.2 Caption-conditioned SAR scene generation

Beyond serving as a benchmark for evaluating visual understanding, the RGBP dataset also enables a novel generative perspective on SAR scenes. Image captioning and image generation form a natural duality: captioning abstracts a visual scene into a compact semantic representation, while generation instantiates such semantics back into a concrete visual realization. This paradigm has driven significant breakthroughs in text-to-image generation, with large-scale image caption datasets such as COCO [42] laying the foundation for modern models like Stable Diffusion [52] and DALL-E [50]. In RGBP, each image is annotated with two factorized captions describing the physical scene and the projected content, respectively, offering similar potential to bridge the gap between high-level human concepts and SAR scenes.

6.1.3 The RGBP dataset with SAR scenes

A second major contribution of our work is addressing the fundamental bottleneck of data scarcity in this domain. Prior to our research, no large-scale, publicly available dataset existed for training and evaluating VLMs on their ability to describe complex SAR scenes.

To solve this, we invested significant effort in the creation of the RGBP dataset. This is a large-scale, specialized resource designed from the ground up for the SAR domain. Its construction was meticulous, involving the capture of 70 distinct real-world physical scenes into which we projected over 180,000 digital images under varied lighting, angle, and occlusion conditions. Crucially, we introduced a unique dual-task ground truth annotation scheme, providing separate, high-quality captions for both the physical scene and the projected content. The creation of the RGBP dataset is a foundational contribution in its own right. It not only provides the essential data needed to train specialized models like ProCap but also offers the wider research community a fair, robust, and challenging benchmark. We believe this dataset will be an invaluable resource, catalyzing further innovation and a deeper understanding of VLM capabilities in mixed-reality environments.

7 Conclusion

In this paper, we addressed a critical gap in the capabilities of modern vision language models. Our investigation, supported by extensive experimental results, confirms that current VLMs often fail when processing complex scenes containing projection information, exhibiting three main drawbacks: (1) an inability to reliably detect the presence of projection; (2) inaccurate descriptions of objects and their spatial arrangement within the physical scene; and (3) a failure to correctly interpret the content of projection.

To solve these problems, we proposed ProCap, a novel two-stage framework that introduces a region-aware understanding paradigm. Our method first employs an automatic segmentation module to disentangle the physical scene from the projected content. It then uses a region-aware retrieval mechanism to gather targeted information for each component before generating separate, accurate descriptions. Recognizing the lack of suitable training data for this task, we also introduced the RGBP dataset, a new, large-scale resource comprising 70 real-world scenes captured with diverse projection content.

Our experiments, conducted using a rigorous dual-task evaluation protocol, show that ProCap significantly outperforms strong baseline models, particularly in the challenging task of describing projected content, while also demonstrating superior generalization and robustness. This work represents a significant step towards enabling VLMs to understand and interact with the increasingly common mixed-reality environments of the modern world.

Acknowledgments

We thank the anonymous reviewers for their valuable and inspiring comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 62302401). Ling was not supported by any fund for this study.

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In *ICCV*, pp. 8947–8956, 2019. 3, 8, 1

[2] AI@Meta. Llama 3 model card, 2024. 2, 4

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., *NeurIPS*, vol. 35, pp. 23716–23736, 2022. 3

[4] R. Asahina, T. Nomoto, T. Yoshida, and Y. Watanabe. Realistic 3d swept-volume display with hidden-surface removal using physical materials. In *IEEE VR*, pp. 113–121, 2021. 2

[5] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. Qwen3-VL technical report, 2025. 2, 3, 6, 7, 4

[6] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL technical report, 2025. 2, 4

[7] A. H. Bermano, M. Billeter, D. Iwai, and A. Grundhöfer. Makeup Lamps: Live augmentation of human faces via projection. *Computer Graphics Forum*, 36(2):311–323, 2017. 2

[8] O. Bimber and R. Raskar. *Spatial Augmented Reality: Merging Real and Virtual Worlds*. A. K. Peters, Ltd., 2005. 1

[9] N. Bitton-Guetta, Y. Bitton, J. Hessel, L. Schmidt, Y. Elovici, G. Stanovsky, and R. Schwartz. Breaking Common Sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. In *ICCV*, pp. 2616–2627, 2023. 3, 8, 1

[10] W. Dai, J. Li, D. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds., *NeurIPS*, vol. 36, pp. 49250–49267, 2023. 3

[11] Y. Deng, H. Ling, and B. Huang. LAPIG: Language guided projector image generation with surface adaptation and stylization. *IEEE TVCG*, 2025. 2

[12] X. Dong, H. Ling, and B. Huang. Adaptive color structured light for calibration and shape reconstruction. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 1240–1249, 2023. 2

[13] Y. Erel, D. Iwai, and A. H. Bermano. Neural projection mapping using reflectance fields. *IEEE TVCG*, 29(11):4339–4349, 2023. 2

[14] Y. Erel, O. Kozlovsky-Mordenfeld, D. Iwai, K. Sato, and A. H. Bermano. Casper DPM: Cascaded perceptual dynamic projection mapping onto hands. In *SIGGRAPH Asia*. ACM, 2024. 2

[15] X. Geng and H. Liu. OpenLLaMA: An open reproduction of LLaMA, 2023. 7, 8, 2, 4

[16] A. Grundhöfer and D. Iwai. Robust, error-tolerant photometric projector compensation. *IEEE TIP*, 24(12):5086–5099, 2015. 2

[17] A. Gupta, P. Dollar, and R. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4, 9

[18] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. 3

[19] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. 3

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[21] B. Huang and H. Ling. CompenNet++: End-to-end full projector compensation. In *ICCV*, 2019. 2

[22] B. Huang and H. Ling. End-to-end projector photometric compensation. In *CVPR*, pp. 6803–6812, 2019. 2

[23] B. Huang and H. Ling. DeProCams: Simultaneous relighting, compensation and shape reconstruction for projector-camera systems. *IEEE TVCG*, 27(5):2725–2735, 2021. 2

[24] B. Huang, T. Sun, and H. Ling. End-to-end full projector compensation. *IEEE TPAMI*, 2022. 2

[25] B. Huang, Y. Tang, S. Ozdemir, and H. Ling. A fast and flexible projector-camera calibration system. *IEEE Transactions on Automation Science and Engineering*, 18(3):1049–1063, 2021. 2

[26] T.-H. Huang, T.-C. Wang, and H. H. Chen. Radiometric compensation of images projected on non-white surfaces by exploiting chromatic adaptation and perceptual anchoring. *IEEE TIP*, 26(1):147–159, 2017. 2

[27] D. Iwai. Projection mapping technologies: A review of current trends and future directions. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 100(3):234–251, 2024. 1

[28] D. Iwai and K. Sato. Limpid desk: see-through access to disorderly desktop in projection-based mixed reality. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, p. 112–115. ACM, 2006. 2

[29] G. Izacard and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880. Association for Computational Linguistics, 2021. 3

[30] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. 3

[31] S. Kagami and K. Hashimoto. Animated Stickies: Fast video projection mapping onto a markerless plane through a direct closed-loop alignment. *IEEE TVCG*, 25(11):3094–3104, 2019. 2

[32] Y. Kageyama, D. Iwai, and K. Sato. Online projector deblurring using a convolutional neural network. *IEEE TVCG*, 28(5):2223–2233, 2022. 2

[33] T. Kaminokado, D. Iwai, and K. Sato. Augmented environment mapping for appearance editing of glossy surfaces. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 55–65, 2019. 2

[34] H. Kusuyama, Y. Kageyama, D. Iwai, and K. Sato. A multi-aperture coaxial projector balancing shadow suppression and deblurring. *IEEE TVCG*, pp. 1–11, 2024. 2

[35] Y. Lee. Qwen2-VL-Finetune, 2024. 6

[36] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, vol. 33, pp. 9459–9474, 2020. 3

[37] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends in Computer Graphics and Vision*, 16(1-2):1–214, 2024. 9

[38] J. Li, Q. Deng, H. Ling, and B. Huang. DPCS: Path tracing-based differentiable projector-camera systems. *IEEE TVCG*, 2025. 2

[39] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3

[40] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

In *ICML*, vol. 162, pp. 12888–12900, 2022. 2

[41] J. Li, D. M. Vo, A. Sugimoto, and H. Nakayama. EVCap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *CVPR*, pp. 13733–13742, 2024. 6, 9

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014. 3, 8, 9, 1

[43] K. Luo, G. Yang, W. Xian, H. Haraldsson, B. Hariharan, and S. Belongie. Stay Positive: Non-negative image synthesis for augmented reality. In *CVPR*, pp. 10050–10060, 2021. 2

[44] T. Nomoto, R. Koishihara, and Y. Watanabe. Realistic dynamic projection mapping using real-time ray tracing. In *SIGGRAPH*. ACM, 2020. 2

[45] T. Nomoto, W. Li, H.-L. Peng, and Y. Watanabe. Dynamic projection mapping with networked multi-projectors based on pixel-parallel intensity control. In *SIGGRAPH Asia*. ACM, 2020. 2

[46] T. Nomoto, W. Li, H.-L. Peng, and Y. Watanabe. Dynamic multi-projection mapping based on parallel intensity control. *IEEE TVCG*, 28(5):2125–2134, 2022. 2

[47] H.-L. Peng, K. Sato, S. Nakagawa, and Y. Watanabe. Perceptually-aligned dynamic facial projection mapping by high-speed face-tracking method and lens-shift co-axial setup. *IEEE TVCG*, 31(10):6824–6838, 2025. 2

[48] H.-L. Peng and Y. Watanabe. High-speed human arm projection mapping with skin deformation. In *SIGGRAPH Asia*. ACM, 2020. 2

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, vol. 139, pp. 8748–8763, 2021. 2

[50] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *ICMR*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021. 9

[51] R. Raskar, J. van Baar, P. Beardsley, T. Willwacher, S. Rao, and C. Forlines. iLamps: geometrically aware and self-configuring projectors. *ACM TOG*, 22(3):809–818, 2003. 2

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022. 9

[53] M. Takeuchi, H. Kusuyama, D. Iwai, and K. Sato. Projection mapping under environmental lighting by replacing room lights with heterogeneous projectors. *IEEE TVCG*, 30(5), 2024. 2

[54] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 6

[55] P. K. A. Vasu, F. Faghri, C.-L. Li, C. Koc, N. True, A. Antony, G. Santhanam, J. Gabriel, P. Grasch, O. Tuzel, and H. Pouransari. FastVLM: Efficient vision encoding for vision language models. In *CVPR*, 2025. 2, 6, 7, 8, 4

[56] T.-J. J. Wang, J. Laaksonen, T. Langer, H. Arponen, and T. E. Bishop. Learning by Hallucinating: Vision-language pre-training with weak supervision. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1073–1083, 2023. 3

[57] Y. Wang, H. Ling, and B. Huang. CompenHR: Efficient full compensation for high-resolution projector. In *IEEE VR*, pp. 135–145, 2023. 2

[58] Y. Wang, H. Ling, and B. Huang. ViComp: Video compensation for projector-camera systems. *IEEE TVCG*, 30(5):2347–2356, 2024. 2

[59] M. Yasui, R. Iwataki, M. Ishikawa, and Y. Watanabe. Projection mapping with a brightly lit surrounding using a mixed light field approach. *IEEE TVCG*, 30(5):2217–2227, 2024. 2

[60] M. Yasunaga, A. Aghajanyan, W. Shi, R. James, J. Leskovec, P. Liang, M. Lewis, L. Zettlemoyer, and W.-T. Yih. Retrieval-augmented multimodal language modeling. In *ICML*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 39755–39769. PMLR, 2023. 3

[61] P. Zhang, G. Zeng, T. Wang, and W. Lu. TinyLlama: An open-source small language model, 2024. 6, 7, 2, 4

[62] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open pre-trained transformer language models, 2022. 6, 7, 2, 4

[63] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. 6, 7, 2, 4

[64] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, Z. Gao, E. Cui, X. Wang, Y. Cao, Y. Liu, X. Wei, H. Zhang, H. Wang, W. Xu, H. Li, J. Wang, N. Deng, S. Li, Y. He, T. Jiang, J. Luo, Y. Wang, C. He, B. Shi, X. Zhang, W. Shao, J. He, Y. Xiong, W. Qu, P. Sun, P. Jiao, H. Lv, L. Wu, K. Zhang, H. Deng, J. Ge, K. Chen, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 2, 4

ProCap: Projection-Aware Captioning for Spatial Augmented Reality

— *Supplementary Materials* —

Zimo Cao[*] Southwest University
Yuchen Deng[†] Southwest University
Haibin Ling[‡] Westlake University
Bingyao Huang[§] Southwest University

A. Introduction

This supplementary material provides extended experimental results and additional qualitative examples to complement main paper. We evaluate ProCap and competing baselines on all 60 trained scenes as well as on 10 challenging untrained scenes. The goal is to offer a more comprehensive view of the model's robustness, generalization ability, and interpretability.

B. RGBP Dataset Details

To further clarify the dataset diversity and coverage, we provide a quantitative summary of surface geometry and lighting conditions in the RGBP dataset. Scenes are explicitly categorized into planar, mildly curved, and highly curved surfaces based on the projection field of view (FOV) covered, with all categories included in both trained and untrained scenes in Tab. 9. Tab. 10 explains lighting conditions in ambient light range and the RGB light panel operating modes. Projection distortions and occlusions naturally emerge from non-planar geometries and surface self-occlusion, particularly in curved scenes. Fig. 6 illustrates lighting, projection deformation, and occlusions are provided to further demonstrate the dataset's coverage.

C. Implementation Details

Our ProCap is implemented in PyTorch and trained for a single epoch using mixed-precision training. We optimize the model with the AdamW optimizer, setting the weight decay to 0.05 and the momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. A cosine learning rate decay strategy is employed with an initial learning rate of 1×10^{-4}. The training process includes 5,000 linear warm-up steps, during which the learning rate is gradually increased from 1×10^{-6} to the initial learning rate. More detailed training information is presented in Tab. 8.

D. Extended Results

Tab. 11 and Tab. 12 report detailed comparisons across COCO [42], nocaps [1], and WHOOPS! [9] subsets. These results collectively confirm that ProCap provides both superior accuracy and generalization in complex SAR scene caption tasks. Several consistent findings emerge:

- In the scene description task across 60 trained scenes, ProCap models achieve superior performance and demonstrate exceptional robustness in distinguishing physical environments from distracting projection content.
- In the projection description task across 60 trained scenes, ProCap demonstrates a decisive advantage over all baselines, particularly in achieving significantly higher CIDEr scores on nocaps and COCO. These results underscore the necessity and robustness of the ProCap approach, which provides more coherent interpretations even when facing the challenging and adversarial content of WHOOPS!.
- In the untrained scenes, ProCap shows limited gains in the scene caption task due to overfitting from the small training set, but fine-tuning on the RGBP dataset enables the model to significantly outperform baselines in the projection caption task. This highlights the RGBP dataset's critical role in enhancing performance within complex SAR environments despite limited scene variety.

Moreover, for scenarios where projected content perfectly aligns with the surface of the projected object in SAR applications, we further evaluated ProCap$_{\text{Vicuna-1.5-7B}}$'s understanding of this PM application. The specific results of these two sets of SAR scenes are shown in Fig. 7. This demonstrates the potential future application value of the ProCap framework in the field of projection mapping.

E. Discussion

The supplementary experiments reinforce several key insights:

- The segmentation-first strategy is essential for disentangled scene understanding in SAR contexts.
- ProCap maintains robust performance under cross-domain shifts, adversarial content, and unseen scenarios.
- The observed performance margins confirm the broader applicability of ProCap as a specialized framework for SAR environments.

Overall, these additional results validate that the proposed segmentation-based, region-aware framework is not only effective but also necessary for advancing visual language understanding in spatial augmented reality.

[*]e-mail: caozimo@email.swu.edu.cn
[†]e-mail: swudyc714@email.swu.edu.cn
[‡]e-mail: linghaibin@westlake.edu.cn
[§]e-mail: bhuang@swu.edu.cn. Corresponding author.

Table 8: Methods on training time and used GPUs.

Method	Training time	GPUs
ProCap $_{\text{TinyLlama-1.1B}}$ [61]	∼4.5h	3 RTX4060Ti
ProCap $_{\text{OpenLLaMA-3B}}$ [15]	∼4.5h	3 RTX4060Ti
ProCap $_{\text{OPT-2.7B}}$ [62]	∼4.3h	3 RTX4060Ti
ProCap $_{\text{Llama-3.3-8B}}$	∼4.2h	1 A100
ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	∼4.2h	1 A100
RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	∼6.0h	1 RTX PRO 6000 Blackwell

Table 9: Detailed configuration of the RGBP dataset in surface curvature variations, measured on projection field of view (FOV) covered.

Surface curvature variations	Trained scenes	Untrained scenes
Planar	01, 02, 03, 04, 05, 06, 07, 08, 09, 21, 25, 28, 29, 37, 45, 47, 60	61, 63, 68, 70
Mildly curved	10, 11, 12, 13, 15, 17, 18, 19, 20, 23, 24, 32, 33, 34, 35, 39, 46, 49, 51, 52, 53, 55, 56	64, 65, 66, 67
Highly curved	14, 16, 22, 26, 27, 30, 31, 36, 38, 40, 41, 42, 43, 44, 48, 50, 54, 57, 58, 59	62, 69

Table 10: Detailed configuration of the RGBP dataset in lighting conditions.

Lighting conditions	Value
Ambient light	approximately 2.9–546.0 lux
RGB light panel	2268 lm / 1235 lm (high / low power modes)

Table 11: Performance comparison on the **60 trained scenes**. Metrics are reported as BLEU@4 (B@4), METEOR (M), CIDEr (C), SPICE (S), where B stands for BLEU, M for METEOR, C for CIDEr, and S for SPICE.

Task	Method	COCO Test				NoCaps val In-domain		Near-domain		Out-domain		Overall		WHOOPS! Test	
		B@4 ↑	M ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene caption	FastVLM-0.5B [55]	4.76	14.25	1.72	9.20	1.87	8.83	1.78	8.93	2.16	10.65	1.60	9.47	1.59	10.17
	FastVLM-1.5B [55]	4.75	14.15	1.92	8.77	1.98	8.44	1.96	8.64	2.33	10.42	1.73	9.16	1.74	9.92
	FastVLM-7B [55]	5.62	15.63	2.31	10.31	2.51	10.13	2.55	10.16	2.87	12.08	2.21	10.79	2.21	11.98
	Qwen2.5-VL-3B-Instruct [6]	7.95	18.56	2.20	11.99	2.28	11.93	2.29	11.86	2.58	12.85	2.00	12.21	1.99	13.04
	Qwen2.5-VL-7B-Instruct [6]	8.34	19.15	2.02	12.45	2.23	12.80	2.24	12.70	2.25	13.03	1.85	12.85	1.87	13.67
	Qwen3-VL-2B-Instruct [5]	8.58	19.18	2.34	12.17	2.42	11.79	2.48	11.82	2.72	12.89	2.12	12.17	2.01	12.83
	Qwen3-VL-4B-Instruct [5]	5.99	18.32	2.38	11.68	2.66	11.80	2.64	11.95	2.64	12.07	2.19	11.94	2.02	12.15
	Qwen3-VL-8B-Instruct [5]	7.19	19.75	2.38	13.03	2.50	13.20	2.50	12.99	2.60	13.24	2.09	13.15	1.97	13.07
	InternVL3-1B [64]	7.57	17.81	1.97	11.64	1.93	11.11	1.92	11.00	2.13	12.48	1.67	11.53	1.67	12.16
	InternVL3-2B [64]	10.39	19.24	3.07	12.37	3.15	12.11	3.11	12.08	3.76	13.41	2.80	12.53	2.74	13.04
	InternVL3-8B [64]	10.00	19.16	3.49	13.00	3.52	13.03	3.57	12.93	3.88	13.75	3.06	13.24	3.01	13.89
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	36.50	29.23	70.27	21.92	51.50	21.54	53.99	23.38	50.03	23.21	35.21	24.21	69.95	21.87
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	35.96	28.89	69.43	22.01	49.69	20.80	55.63	22.64	49.30	23.17	37.14	22.75	69.46	22.08
	ProCap $_{\text{Llama-3.3-8B}}$	33.36	28.82	35.64	22.24	35.16	22.19	35.46	22.21	35.07	22.20	35.21	22.20	35.49	22.14
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	34.15	29.13	36.17	22.75	36.98	22.94	36.22	22.84	36.07	22.85	36.39	22.88	36.53	22.70
	ProCap $_{\text{OPT-2.7B}}$ [62]	35.95	28.75	28.01	24.34	28.44	24.44	28.34	24.46	28.46	24.48	27.63	24.45	28.19	24.11
	RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	36.38	29.18	37.81	23.22	37.41	23.47	37.80	23.48	37.32	23.54	37.48	23.49	37.70	23.59
Projection caption	FastVLM-0.5B [55]	4.58	14.18	7.17	7.43	6.91	6.39	6.93	6.32	3.89	4.35	6.16	5.68	5.54	5.85
	FastVLM-1.5B [55]	4.75	14.02	7.21	7.31	6.82	6.79	6.20	6.36	3.64	4.48	5.72	5.88	4.14	6.09
	FastVLM-7B [55]	5.26	14.15	7.65	7.33	7.11	6.77	7.30	6.88	5.91	4.69	7.01	6.12	6.25	6.77
	Qwen2.5-VL-3B-Instruct [6]	6.85	17.33	8.41	10.52	10.67	9.00	12.33	9.11	9.04	7.48	10.79	8.53	7.37	8.92
	Qwen2.5-VL-7B-Instruct [6]	7.32	18.19	14.07	11.43	16.52	9.67	18.14	9.54	15.95	8.53	17.11	9.25	12.15	10.27
	Qwen3-VL-2B-Instruct [5]	7.97	18.80	10.73	12.47	14.41	11.11	16.47	10.93	13.39	9.12	14.85	10.38	9.82	11.05
	Qwen3-VL-4B-Instruct [5]	5.85	18.18	10.59	11.72	14.64	11.14	17.06	10.90	13.25	9.37	14.93	10.47	10.35	10.91
	Qwen3-VL-8B-Instruct [5]	6.10	17.83	11.56	11.63	15.38	10.31	17.64	10.66	13.92	9.07	15.57	10.01	11.19	10.97
	InternVL3-1B [64]	6.63	17.04	8.02	10.33	10.95	8.91	12.35	8.79	8.49	6.58	10.78	8.10	6.89	9.12
	InternVL3-2B [64]	11.06	18.81	37.71	11.67	37.68	9.50	42.89	9.36	36.14	6.93	39.66	8.60	34.43	11.02
	InternVL3-8B [64]	11.14	20.25	38.68	13.67	36.94	10.99	42.05	11.12	37.89	8.75	39.31	10.29	34.55	12.27
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	15.68	16.61	54.37	9.75	39.62	5.74	29.77	4.79	17.69	3.64	30.45	4.80	11.88	3.62
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	24.18	20.59	76.98	13.58	57.61	8.43	42.73	6.82	24.10	4.57	42.06	6.66	19.37	5.43
	ProCap $_{\text{Llama-3.3-8B}}$	25.60	21.18	79.10	14.10	52.57	8.18	37.16	6.47	24.17	4.89	38.57	6.51	20.09	5.71
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	24.58	21.05	78.99	13.83	55.19	8.32	39.75	6.68	22.85	4.75	39.93	6.59	22.33	6.02
	ProCap $_{\text{OPT-2.7B}}$ [62]	18.05	16.83	57.72	10.21	46.48	6.91	30.40	5.41	13.03	3.19	30.58	5.17	15.04	4.40
	RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	36.37	27.75	127.58	21.14	99.52	13.17	107.20	13.74	98.00	12.68	102.67	13.19	80.46	16.07

Figure 6: This figure shows the realistic capture effect rendering of two sets of test digital images projected in all 60 training scenes.

Table 12: Performance comparison on the **5 untrained scenes**. Metrics are reported as BLEU@4 (B@4), METEOR (M), CIDEr (C), SPICE (S), where B stands for BLEU, M for METEOR, C for CIDEr, and S for SPICE.

Task	Method	COCO Test				NoCaps val In-domain		Near-domain		Out-domain		Overall		WHOOPS! Test	
		B@4 ↑	M ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Scene caption	FastVLM-0.5B [55]	2.16	10.98	1.44	5.42	1.42	5.20	1.50	4.96	1.90	6.62	1.32	5.58	1.02	6.78
	FastVLM-1.5B [55]	1.88	10.90	0.98	5.36	1.14	5.14	1.14	5.98	1.48	8.16	1.04	6.44	0.84	8.50
	FastVLM-7B [55]	2.10	11.64	1.24	6.00	1.60	6.30	1.48	6.02	1.88	7.80	1.38	6.72	1.34	8.66
	Qwen2.5-VL-3B-Instruct [6]	4.40	14.08	2.14	7.50	2.36	7.50	2.46	7.30	2.52	8.58	2.10	7.78	1.82	8.94
	Qwen2.5-VL-7B-Instruct [6]	5.50	14.88	2.12	8.38	2.32	8.32	2.36	8.46	2.60	9.26	2.04	8.68	1.70	8.96
	Qwen3-VL-2B-Instruct [5]	4.48	15.42	2.36	9.32	2.56	8.66	2.66	8.64	2.66	9.24	2.18	8.86	1.86	8.94
	Qwen3-VL-4B-Instruct [5]	5.06	16.60	3.44	9.08	4.02	8.64	3.98	8.72	4.12	9.26	3.42	8.88	2.98	9.04
	Qwen3-VL-8B-Instruct [5]	5.32	17.32	3.68	12.26	4.12	11.88	3.96	11.86	4.16	12.32	3.52	12.00	3.08	12.46
	InternVL3-1B [64]	4.34	14.74	1.26	8.20	1.28	7.76	1.28	7.56	1.76	9.68	1.20	8.34	1.06	8.80
	InternVL3-2B [64]	5.90	15.18	2.22	9.24	2.54	8.84	2.58	8.98	3.62	10.82	2.40	9.54	2.12	10.40
	InternVL3-8B [64]	4.84	15.82	2.38	10.14	2.70	9.94	2.80	9.86	3.02	10.72	2.34	10.16	2.24	10.98
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	2.48	11.38	5.02	5.53	4.94	8.34	4.74	8.34	6.08	8.50	4.06	8.34	7.48	5.15
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	1.50	11.73	3.78	7.77	4.88	7.16	4.96	7.28	5.56	7.14	4.02	7.10	5.73	6.60
	ProCap $_{\text{Llama-3.3-8B}}$ [2]	2.22	12.10	2.92	7.44	3.20	6.94	3.02	6.68	3.32	7.12	2.64	6.92	2.46	6.90
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	1.90	11.66	2.94	7.50	2.80	7.46	2.76	7.18	3.04	7.26	2.34	7.30	2.22	7.12
	ProCap $_{\text{OPT-2.7B}}$ [62]	4.10	14.46	5.80	9.44	5.78	9.06	5.62	9.24	5.50	9.06	4.82	9.12	4.68	8.90
	RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	5.64	13.64	4.82	9.40	5.76	8.84	6.04	8.96	5.92	8.80	4.96	8.86	4.00	9.42
Projection caption	FastVLM-0.5B [55]	6.28	16.28	11.54	9.38	8.40	7.60	10.48	7.92	6.42	5.66	8.94	7.02	8.72	8.06
	FastVLM-1.5B [55]	6.20	15.38	10.00	8.92	7.58	7.84	7.48	7.86	4.60	6.02	6.84	7.26	4.88	7.70
	FastVLM-7B [55]	6.44	16.36	9.70	9.38	8.62	8.50	9.70	9.20	8.80	6.62	9.44	8.14	7.84	9.08
	Qwen2.5-VL-3B-Instruct [6]	9.18	20.52	11.88	14.24	14.52	11.64	16.80	11.48	12.18	10.30	14.66	11.14	9.52	11.78
	Qwen2.5-VL-7B-Instruct [6]	9.32	21.16	19.52	14.60	20.56	11.64	24.60	11.96	23.12	11.34	23.00	11.62	16.28	12.70
	Qwen3-VL-2B-Instruct [5]	10.40	22.18	14.10	16.14	17.96	13.34	21.50	13.24	19.16	12.04	19.60	12.84	13.04	13.78
	Qwen3-VL-4B-Instruct [5]	6.82	20.44	12.86	14.32	17.46	12.90	20.34	13.16	17.74	11.72	18.46	12.58	12.72	12.88
	Qwen3-VL-8B-Instruct [5]	7.52	20.58	14.34	14.68	18.42	12.30	21.46	13.04	18.06	11.64	19.20	12.34	13.90	13.10
	InternVL3-1B [64]	8.88	19.72	10.56	13.64	14.02	11.00	15.86	10.88	11.86	8.58	14.12	10.16	9.28	11.92
	InternVL3-2B [64]	13.78	21.22	46.94	14.24	46.24	11.24	52.88	11.00	46.54	8.26	49.14	10.18	41.40	13.34
	InternVL3-8B [64]	13.10	22.90	45.84	16.80	42.98	12.82	52.10	13.38	46.80	11.00	47.74	12.42	43.10	14.82
	ProCap $_{\text{TinyLlama-1.1B}}$ [61]	13.13	17.92	46.90	12.60	42.32	6.10	33.64	5.26	19.70	3.96	32.36	5.10	15.23	4.57
	ProCap $_{\text{OpenLLaMA-3B}}$ [15]	19.90	21.22	66.03	14.90	58.22	8.66	48.00	7.34	27.10	5.20	44.72	7.06	24.65	6.80
	ProCap $_{\text{Llama-3.3-8B}}$ [2]	27.82	21.98	85.02	15.02	56.68	8.56	38.20	6.62	27.02	5.42	41.12	6.88	21.00	5.92
	ProCap $_{\text{Vicuna-1.5-7B}}$ [63]	26.20	22.08	86.26	14.88	58.70	8.84	43.46	7.18	27.32	5.30	43.94	7.10	24.44	6.60
	ProCap $_{\text{OPT-2.7B}}$ [62]	19.70	17.78	63.40	10.96	51.70	7.48	32.22	5.78	16.50	3.82	34.18	5.68	16.72	4.82
	RGBP $_{\text{Qwen3-VL-8B-instruct}}$ [5]	38.18	28.90	136.60	22.24	104.50	13.54	113.76	14.36	103.94	13.30	108.66	13.76	85.82	17.02

Table 13: Performance comparison on **the first 5 untrained scenes**. Metrics are reported as BLEU@4 (B@4), METEOR (M), CIDEr (C), SPICE (S). Results are averaged on the 5 scenes, and the best results are in **bold**.

Task	Method	COCO Test				nocaps val In-domain		Near-domain		Out-domain		Overall		WHOOPS! Test	
		B@4 ↑	M ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑	C ↑	S ↑
Projection	ProCap $_{\text{Vicuna-1.5-7B, trained scene}}$	24.58	21.05	78.99	13.83	55.19	8.32	39.75	6.68	22.85	4.75	39.93	6.59	22.33	6.02
	ProCap $_{\text{Vicuna-1.5-7B, untrained scene}}$	**26.20**	22.08	86.26	14.88	58.70	8.84	43.46	7.18	**27.32**	**5.30**	**43.94**	**7.10**	24.44	6.60
	ProCap $_{\text{Vicuna-1.5-7B, projection}}$	**26.20**	**23.00**	**89.80**	**15.60**	**63.6**	**9.40**	**45.60**	**7.90**	18.70	3.90	43.50	**7.10**	**27.60**	**7.40**

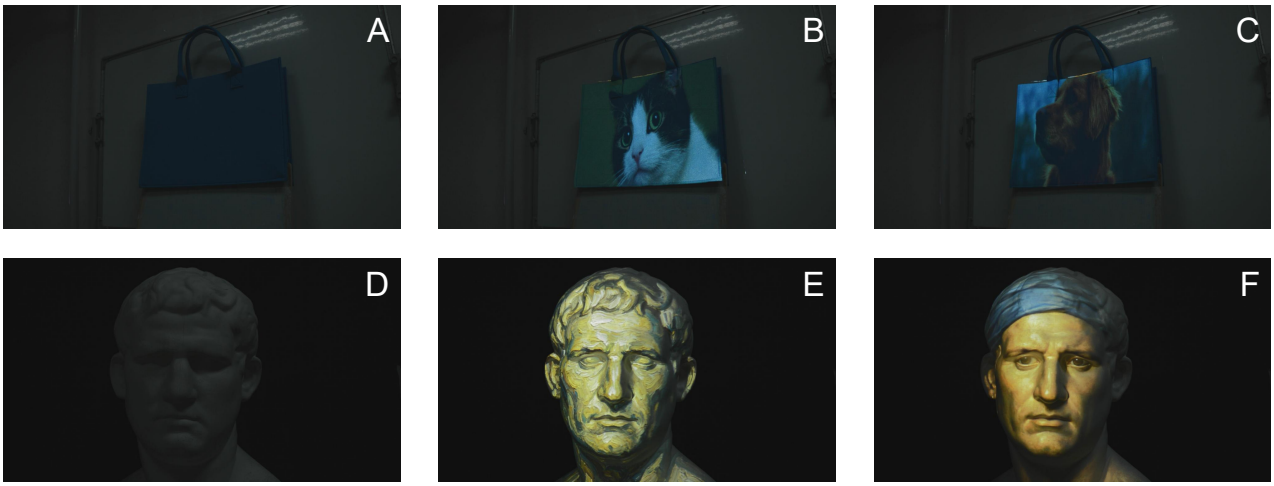

Figure 7: Examples of projection mapping based on TOSHIBA TDP-T100C projector (1024 $\times$ 768 resolution) and Nikon D3200 camera (1280 $\times$ 720 resolution) using ProCap $_{\text{Vicuna-1.5-7B}}$. We highlight incorrect captioned objects in red and correct ones in blue. (A) and (D) is the physical scene w/o projected content. (B) scene caption: *"A blue carrier bag, with its white handles and dark blue panels, is in front of the white wall."*, projection caption: *"A black and white cat sitting on a wooden table."*. (C) scene caption: *"A blue paper bag with white handles is in front of a white wall."*, projection caption: *"A brown and white dog wearing a red collar."*. (E) scene caption: *"A view of a bust of a man with curly hair."*, projection caption: *" A man is wearing a suit and tie."*. (F) scene caption: *"There is a white statue of a man's head with curly hair."*, projection caption: *" A man is wearing a suit and tie."*. Note that the model outputs are truncated by the max tokens limit.