When Big Models Train Small Ones: Label-Free Model Parity Alignment for Efficient Visual Question Answering using Small VLMs

Abhirama Subramanyam Penamakuri*, **Navlika Singh***, **Piyush Arora*** **and Anand Mishra**
Indian Institute of Technology Jodhpur
{penamakuri.1,singh.119,arora.8,mishra}@iitj.ac.in
https://github.com/vl2g/MPA

Abstract

Large Vision-Language Models (L-VLMs) have demonstrated remarkable performance in various vision and language tasks, including visual question answering (VQA). However, their high computational cost makes them impractical for resource-constrained settings and inference-heavy applications. In contrast, Small Vision-Language Models (S-VLMs) offer efficiency but suffer from a significant performance gap compared to their larger counterparts. In this work, we introduce the Model Parity Aligner (MPA), a novel framework designed to systematically improve S-VLMs by leveraging unlabeled images and effective knowledge transfer from L-VLMs. Instead of traditional knowledge distillation methods that rely on labeled training data, MPA employs a strategic parity-based approach that precisely identifies the knowledge disparities between S-VLMs and L-VLMs, and optimizes training by targeting only these disparities. We conduct extensive experiments on four diverse VQA benchmarks, namely TextVQA, ST-VQA, ChartQA, and OKVQA, each of which required specialized reasoning capabilities such as text recognition, chart interpretation, and commonsense and factual understanding. Our results demonstrate that MPA consistently enhances the performance of S-VLM on all benchmarks, reducing the performance gap while maintaining computational efficiency. We make our code publicly available.

1 Introduction

Large vision and language models (L-VLMs) have recently made remarkable progress on various vision and language tasks, including visual question answering (VQA) (Liu et al., 2024; Dai et al., 2024; Li et al., 2023; Zhu et al., 2023; Ye et al., 2023; Wang et al., 2024; Chen et al., 2024; Ghosh et al., 2024). This makes them a de facto first choice for the VQA task on a new data set that does not have labeled training samples. However, L-VLMs may not be the most practical choice in resource-constrained settings and especially for inference-heavy tasks such as VQA, due to their high computational requirements and latency. In contrast, smaller vision and language models (S-VLMs) are more efficient but fall significantly short in performance, as shown in Figure 1. This raises a critical question: *Can we improve* S-VLM*s by a relevant and effective knowledge transfer from* L-VLM*s?*

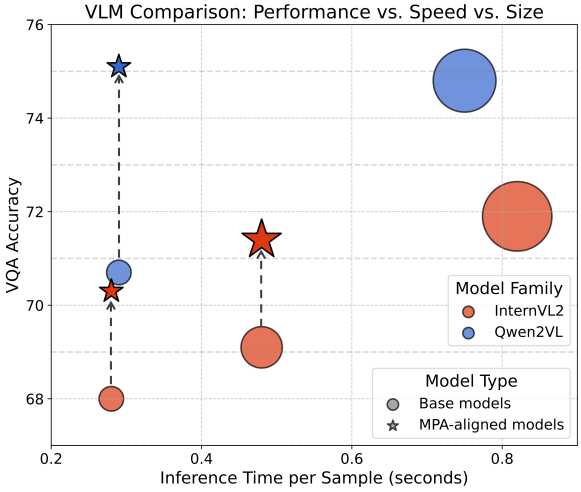


Figure 1: Small models often struggle to match the performance of their larger counterparts. We show model sizes using circle with radius proportional to the parameter count, and their respective inference time and VQA accuracies in X and Y-axis, respectively on one of the datasets used in this paper (Singh et al., 2019). Proposed MPA significantly enhances VQA accuracy for five S-VLMs across four datasets. **(Best viewed in color)**.

Several techniques have been explored to transfer knowledge from large neural models to smaller ones such as: (i) knowledge distillation (KD) (Hinton et al., 2015; Sanh et al., 2019; Gu et al., 2024; Ko et al., 2024; Xu et al., 2024; Shu et al., 2024; Cai et al., 2024) trains a small model (student) to mimic a large model (teacher) by learning from its

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31633–31650
November 4-9, 2025

soft labels or intermediate representations. However, KD typically relies on labeled training data, which may not always be available, and effectively distilling multimodal knowledge remains a challenge due to the complex interplay between vision and language features. (ii) Adapter-based methods (Houlsby et al., 2019; Hu et al., 2022; Liu et al., 2022; Dettmers et al., 2023) introduce lightweight trainable layers into large models to enable efficient fine-tuning. Although these methods reduce training costs, they still require access to large models during inference, limiting their practical advantages in resource-constrained environments. (iii) Self-supervised learning and pseudo-labeling (Chen et al., 2013; Veit et al., 2017; Radosavovic et al., 2018; Xie et al., 2020; Khan et al., 2023) provide an alternative by leveraging unlabeled data to generate training signals. However, naïve pseudo-labeling often propagates noisy predictions, reducing overall effectiveness. Moreover, the challenge of systematically transferring knowledge from large to small vision-language models using pseudo-labeling remains largely under explored. Addressing this gap is crucial for making smaller models more capable without the high computational cost of large models for inference.

To fill the aforementioned gaps, we introduce the Model Parity Aligner (MPA) – a framework that enables effective knowledge transfer from L-VLM to S-VLM using only unlabeled images. Instead of relying on traditional knowledge distillation or fine-tuning, MPA utilizes large model-guided pseudo-labeling with quality assessment. MPA accurately identifies and addresses the knowledge gaps between S-VLM and L-VLM, ensuring that small models learn from high-confidence predictions while minimizing error propagation. By leveraging the strong reasoning capabilities of large VLMs to create high-quality supervision signals through systematic parity assessment, MPA efficiently addresses performance gaps while maintaining computational efficiency.

We conducted extensive experiments and ablation studies to evaluate the effectiveness of the MPA. Specifically, we used four public datasets – TextVQA (Singh et al., 2019), ST-VQA (Biten et al., 2019), ChartQA (Masry et al., 2022), and OKVQA (Marino et al., 2019). These datasets require additional capabilities such as visual text understanding, chart interpretation, and world knowledge integration, making them well-suited to test the robustness of MPA. We experimented with ten combinations of L-VLM and S-VLM pairs, demonstrating that MPA consistently improves S-VLM performance across all benchmarks, highlighting its effectiveness in knowledge transfer.

Contributions: (i) We propose a Model Parity Aligner (MPA) – an effective approach that empowers small VLMs and improve their visual question-answering performance using only unlabeled images, eliminating the need for expensive labeled datasets. (ii) MPA employs a novel parity-based training paradigm, leveraging the L-VLM to generate pseudo-labels for unlabeled images while identifying and targeting specific knowledge gaps between S-VLM and L-VLM. This strategy ensures reliable supervision, minimizes noise, and maximizes relevant knowledge transfer. (iii) MPA achieves consistent improvement across four diverse VQA benchmarks. Furthermore, our findings indicate that MPA not only improves VQA performance, but also enables S-VLM to benefit from closed-source L-VLMs and enhances its core capabilities beyond VQA, such as text recognition and text-aware captioning.

2 Related Work

Small and Large VLMs: Following the success of large language models (LLMs) (Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023; Penedo et al., 2024; Dubey et al., 2024; Yang et al., 2024) across NLP tasks, vision and language models (VLMs) (Liu et al., 2024; Dai et al., 2024; Zhu et al., 2023; Ye et al., 2023; Chen et al., 2024; Wang et al., 2024; Zhou et al., 2024; Marafioti, 2024) have been developed that process both visual and textual data. Although state-of-the-art VLMs achieve impressive zero-shot performance, their growing parameter count impose significant constraints on computational efficiency, accessibility, and deployment costs. This trade-off between efficiency and capacity requires the development of smaller VLMs (Zhou et al., 2024; Shao et al., 2024; Marafioti, 2024) that maintain competitive performance with reduced computational demands (Lu et al., 2024). The key approach to developing S-VLMs from L-VLMs involves substituting the internal LLM with lightweight alternatives (Team et al., 2024; Abdin et al., 2024; Zhang et al., 2024; HuggingFaceTB, 2023). Inspired by the literature on LLM (Lu et al., 2024), we follow a parameter-based taxonomy where VLMs with $\leq$ 5B parameters are classified as S-VLMs, while those that

exceed this threshold are L-VLMs. For context, a small 4B-parameter VLM constitutes just 0.2% of the estimated 1.8T parameters of GPT-4.

Knowledge Distillation: It transfers knowledge from large teacher models to smaller student models using KL-divergence over soft logits (Hinton et al., 2015) or feature representations (Wang et al., 2021; Xu et al., 2020; Sanh et al., 2019). With LLMs adhering to scaling laws, their distillation has gained significant interest. Recent methods for LLMs (Gu et al., 2024; Ko et al., 2024) and L-VLMs (Shu et al., 2024; Xu et al., 2024; Cai et al., 2024) explore KL-Divergence variants, while others (Hsieh et al., 2023; Tian et al., 2024; Ranaldi and Freitas, 2024) distill reasoning via LLM-generated Chain-of-Thought rationales. In contrast to standard KD, which distills over the labeled dataset, our method identifies and supervises only the samples that represent knowledge gaps between the student and teacher. This targeted strategy enables efficient, model-agnostic training using only input-output access to the teacher–including closed-source L-VLMs.

Data Augmentation for VQA: Vision and language tasks such as VQA have traditionally been benefited by data augmentation, and visual question generation becomes a natural choice to generate augmented data (Fan et al., 2018; Jain et al., 2017; Krishna et al., 2019; Mostafazadeh et al., 2016; Wang et al., 2022; Jahagirdar et al., 2021; Zhang et al., 2017; Vedd et al., 2022). Although few methods (Chen et al., 2022; Kant et al., 2021; Kil et al., 2021; Khan et al., 2023) augment the data-scarce VQA datasets to improve performance, other methods (Banerjee et al., 2021; Changpinyo et al., 2022) leverage large-scale image-caption datasets to generate noisy VQA labels and use them as VQA foundational data. Distinctively different from these lines of work, we employ L-VLMs to pseudo-label unlabeled images with a quality check to discard noise, ensuring minimal yet effective annotations for targeted improvements of S-VLMs.

3 Model Parity Aligner (MPA)

Given a task $\mathcal{T}$ and a set of unlabeled images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, our goal is to empower small vision language models (S-VLMs) with task-specific capabilities and improve their performance on the task $\mathcal{T}$. In this work, we restrict ourselves to the VQA task and experiment with various variants of VQA that require interpretation of visual text, chart, and external knowledge. Inspired by the standard machine learning lifecycle (Stodden, 2020), our proposed Model Parity Aligner (MPA) framework follows a systematic approach to achieve this goal. The process begins with automatically annotating unlabeled images $\mathcal{I}$ for task $\mathcal{T}$ using the Pseudo Annotator module discussed in Section 3.1, followed by strategic data selection with automatic quality assessment of the annotations using the Parity Identification module discussed in Section 3.2. This automatically curated and cleaned data is then utilized to fine-tune the S-VLM model using the Parity Leveler module discussed in Section 3.3. The workflow of our proposed MPA framework, which includes its three interconnected modules, is illustrated in Figure 2 and described in Algorithm 1.

Algorithm 1 Model Parity Aligner (MPA)

Input: Large Vision Language Model (L-VLM) parameterized by ϕ; Small Vision-Language Model (S-VLM) parameterized by θ; unlabeled images: $\mathcal{I} = \{I_1, I_2, \cdots, I_N\}$; **task**: $\mathcal{T}$.

Output: Enhanced S-VLM with updated parameters ($\hat{\theta}$).

1: $\mathcal{D}_{PA}^{\mathcal{T}} \leftarrow$ **PA**(L-VLM$_\phi$, $\mathcal{I}$, $\mathcal{T}$) ▷ PA - Pseudo Annotator, $\mathcal{D}_{PA}^{\mathcal{T}}$: pseudo-annotated data

2: $\mathcal{D}_{PI}^{\mathcal{T}} \leftarrow$ **PI**(L-VLM$_\phi$, S-VLM$_\theta$, $\mathcal{D}_{PA}^{\mathcal{T}}$) ▷ PI - Parity Identifier, $\mathcal{D}_{PI}^{\mathcal{T}}$: parity dataset

3: S-VLM$_{\hat{\theta}} \leftarrow$ **PL**(S-VLM$_\theta$, $\mathcal{D}_{PI}^{\mathcal{T}}$) ▷ PL: Parity Leveler

4: **return** S-VLM$_{\hat{\theta}}$

The proposed Model Parity Aligner (MPA) consists of three main modules: (a) Pseudo Annotator (PA), (b) Parity Identifier (PI), (c) Parity Leveler (PL). These modules work together to systematically enrich S-VLMs. The MPA takes S-VLM$_\theta$, L-VLM$_\phi$, a set of unlabeled images $\mathcal{I}$ and the task $\mathcal{T}$ as inputs and returns an enhanced S-VLM$_{\hat{\theta}}$ where ϕ, θ, $\hat{\theta}$ are the parameters of L-VLM, S-VLM, updated S-VLM, respectively. It should be noted here that $|\hat{\theta}| = |\theta| \ll |\phi|$ where $|\cdot|$ denotes the size of the model. Next, we provide an in-depth overview of each module.

3.1 Pseudo Annotator (PA)

This module which is described in Algorithm 2 is responsible for obtaining pseudo-annotation for unlabeled images $\mathcal{I}$. We employ an L-VLM to generate annotations for the unlabeled images for the task $\mathcal{T}$. In this work, we experimented with two L-VLMs. Since we have only access to unlabeled

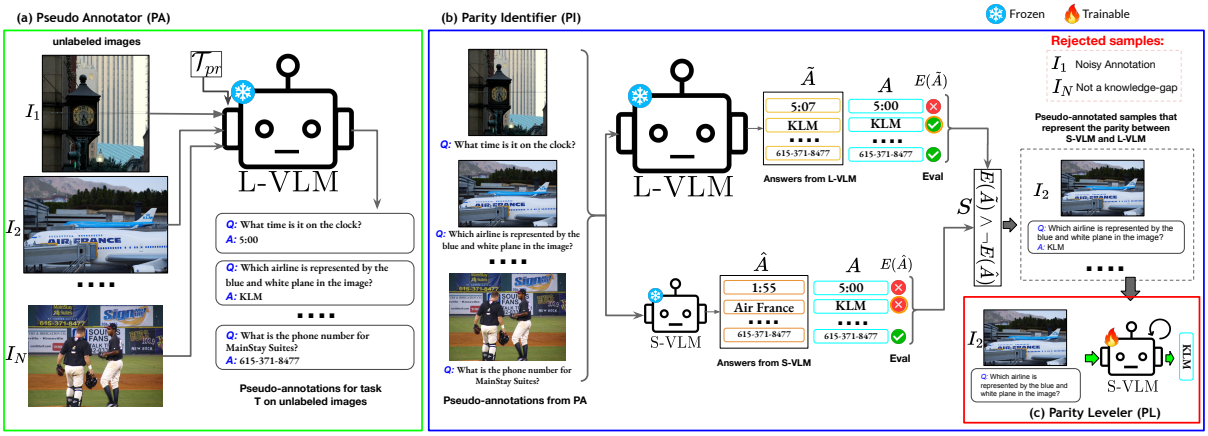


Figure 2: **Overview of the proposed MPA framework.** It consists of three modules, namely (a) Pseudo Annotator (Section 3.1), (b) Parity Identifier (Section 3.2), and (c) Parity Leveler (Section 3.3). Given a set of unlabeled images $\mathcal{I}$ and task $\mathcal{T}$, MPA begins with automatically annotating the unlabeled images, followed by strategic data selection that targets knowledge gaps of S-VLM with the L-VLM, while accounting for annotation quality. This selection process identifies parity, capturing instances where the L-VLM answers correctly while the S-VLM fails. Finally, PL updates the S-VLM's parameters on the obtained parity subset. **(Best viewed in color)**.

Algorithm 2 Pseudo Annotator (PA)

Input: Large Vision Language Model (L-VLM) parameterized by ϕ; unlabeled images: $\mathcal{I} = \{I_i\}_{i=1}^{N}$; task prompt: $\mathcal{T}_{pr}$.
Output: pseudo-annotated images for the **task** $\mathcal{T}$: $\mathcal{D}_{PA}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$.

$\mathcal{D}_{PA}^{\mathcal{T}} \leftarrow [\ ]$
for I_i in $\mathcal{I}$ **do**
  $(Q, A)_i \leftarrow \text{L-VLM}_{\phi}(\mathcal{T}_{pr}, I_i)$
  $\mathcal{D}_{PA}^{\mathcal{T}}$.append($(I_i, Q_i, A_i)$) ▷ Triplet: (I_i, Q_i, A_i) is considered as one pseudo-annotated sample for task $\mathcal{T}$.
end for
return $\mathcal{D}_{PA}^{\mathcal{T}}$

images, we ask L-VLM to generate task-specific visual question and answer pairs. The generation of visual questions (VQG) has been shown to improve the visio-lingual abilities of a vision and language model (Kafle et al., 2017; Chen et al., 2022). In this work, we additionally ask L-VLM to generate the corresponding answer.

To be precise, L-VLM is prompted with a task-specific prompt $\mathcal{T}_{pr}$ to create task-specific question-answer pairs[1] $(Q, A)_i$ for each image I_i within $\mathcal{I}$, where $i \in \{1, \cdots, N\}$. The module produces the pseudo-annotated dataset $\mathcal{D}_{PA}^{\mathcal{T}}$ for task $\mathcal{T}$: $\{(I_i, Q_i, A_i)\}_{i=1}^{N}$, with each triplet (I_i, Q_i, A_i) representing an annotated sample for task $\mathcal{T}$. The L-VLM-driven automated annotation presents challenges, e.g., (i) noisy annotations and (ii) hallucinated content necessitating careful quality validations. Our proposed PI module, described next, inherently accounts for quality validations and minimizes such noisy annotations, while sampling for parity samples.

[1]For example, in the case of ChartQA, $\mathcal{T}_{pr}$ instructs the model to focus on reasoning over charts, including trend analysis and numerical interpretation. Similarly, for TextVQA, the prompt emphasizes reading and comprehending scene text to formulate relevant questions and answers. This ensures that the generated QA pairs align with the specific reasoning challenges posed by each task.

3.2 Parity Identifier (PI)

This module capitalizes on the existing capabilities of S-VLM while isolating its knowledge gaps relative to L-VLM. Rather than following conventional approaches (Chen et al., 2022; Khan et al., 2023; Changpinyo et al., 2022) of using all pseudo-annotated data for training, we implement a more targeted methodology to identify specific knowledge disparities between models. We evaluated both L-VLM and S-VLM in zero-shot settings by presenting each model with image-question pairs (I_i, Q_i) from the PA-annotated dataset $\mathcal{D}_{PA}^{\mathcal{T}}$. The respective answers - $\tilde{A}_i$ from L-VLM and $\hat{A}_i$ from S-VLM—are then compared against the pseudo annotation A_i using the following expression.

$$E(X) = \begin{cases} 1, & \text{if } X = A, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } X \in \{\tilde{A}, \hat{A}\}. \tag{1}$$

Further, we select samples that satisfy the following Boolean condition S.

$$S\big((I, Q, A)\big) = \begin{cases} 1, & \text{if } E(\tilde{A}) \wedge \neg E(\hat{A}), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Algorithm 3 Parity Identifier (PI)

Input: Large Vision Language Model (L-VLM) parameterized by ϕ; Small Vision Language Model (S-VLM) parameterized by θ; pseudo-annotated data: $\mathcal{D}_{PA}^{\mathcal{T}}$.
Output: Parity (Knowledge gap) between L-VLM and S-VLM: $\mathcal{D}_{PI}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{K}, K \ll N$.

$\mathcal{D}_{PI}^{\mathcal{T}} = [\ ]$
for (I_i, Q_i, A_i) in $\mathcal{D}_{PA}^{\mathcal{T}}$ **do**
  $\tilde{A}_i \leftarrow \text{L-VLM}_{\phi}(I_i, Q_i)$
  $\hat{A}_i \leftarrow \text{S-VLM}_{\theta}(I_i, Q_i)$
  if $\tilde{A}_i == A_i$ **and** $\hat{A}_i \neq A_i$ **then** ▷ Eq. 1 & Eq. 2
    $\mathcal{D}_{PI}^{\mathcal{T}}$.append($(I_i, Q_i, A_i)$) ▷ Satisfies Eq. 2 criteria
  else
    continue
  end if
end for
return $\mathcal{D}_{PI}^{\mathcal{T}}$

Here, Boolean condition S selects an annotated triplet (I_i, Q_i, A_i) if $\tilde{A}_i$ correctly matches A_i while $\hat{A}_i$ does not, thereby precisely identifying the knowledge gap between the models where S-VLM requires improvement. In other words, S selects those samples where L-VLM answers correctly, while S-VLM answer is incorrect, assuming the pseudo-annotated answer as ground truth. This methodology inherently performs quality verification by leveraging L-VLM's superior answering capabilities, as these models are primarily instruction-tuned for answering rather than annotating. By selecting only instances where L-VLM demonstrates consistency between its annotation and answering phases, PI module effectively filters out noisy or hallucinated annotations. The resulting parity subset $\mathcal{D}_{PI}^{\mathcal{T}} : \{(I_i, Q_i, A_i)\}_{i=1}^{K}$ with $K \ll N$, constitutes highly efficient samples focused exclusively on the specific knowledge deficiencies of S-VLM. This targeted approach eliminates the need to train on potentially problematic samples or the entire annotation set, optimizing both training efficiency and model performance. This module is detailed in Algorithm 3.

3.3 Parity Leveler (PL)

This module fine-tunes S-VLM on the parity (knowledge gap) samples identified by the PI module for L number of iterations. We feed each sample $\{I, Q\}_i$ from $\mathcal{D}_{PI}^{\mathcal{T}}$, within an instruction prompt template

Algorithm 4 Parity Leveler (PL)

Input: Small Vision-Language Model (S-VLM) parameterized by θ; parity set: $\mathcal{D}_{PI}^{\mathcal{T}} = \{(I_i, Q_i, A_i)\}_{i=1}^{K}$
Output: Enhanced S-VLM with updated parameters $\hat{\theta}$

for iter $= 1$ to L **do** ▷ L: total no. of iterations
  for $\{(I_i, Q_i, A_i)\}_{i=1}^{b}$ in $\mathcal{D}_{PI}^{\mathcal{T}}$ **do** ▷ b: batch size
    $\{\hat{A}_i\}_{i=1}^{b} \leftarrow \text{S-VLM}_{\theta}(\{(I_i, Q_i)\}_{i=1}^{b})$
    Compute $\mathcal{L}_{gen}(\{\hat{A}_i, A_i\}_{i=1}^{b})$ ▷ Answer generation loss
    Update θ using $\mathcal{L}_{gen}$ ▷ Gradient descent
  end for
end for
return S-VLM$_{\hat{\theta}}$

to S-VLM to generate the accurate answer A_i to the visual question Q_i on the image I_i. S-VLM learns $P(A_i|Q_i, I_i)$ by modeling the task as a text generation problem, auto-regressively generating the tokens in the answer.

$$\mathcal{L}_{gen}(\theta) = -\frac{1}{b}\sum_{i=1}^{b}\left[\sum_{t=1}^{m}\log P_{\theta}(A_{i_t}|A_{i<t}, \{I_i, Q_i\})\right] \quad (3)$$

Once all answer tokens $A_{i_{1:m}}$ are obtained, we optimize the model using the generation loss $\mathcal{L}_{gen}$, defined over the minibatches of size b samples (Eq.3) which is minimized via stochastic gradient descent. Note that L-VLM parameters ϕ remain frozen throughout MPA. For an algorithmic description of this module, refer to Algorithm 4.

4 Experiments and Results

Datasets. We evaluate our approach on four widely-used public VQA benchmarks, namely, TextVQA (Singh et al., 2019), ST-VQA (Biten et al., 2019), ChartQA (Masry et al., 2022), and OKVQA (Marino et al., 2019). These datasets are relevant to MPA because they introduce diverse reasoning challenges, such as text, chart, external world understanding beyond traditional VQA (Antol et al., 2015), making them strong benchmarks for evaluating gains in S-VLM. More details on these datasets are in Appendix B. Further, as MPA is primarily designed for label-free training, we exclude all question-answer annotations from the training splits of each dataset during evaluation.

S-VLMs and L-VLMs used. Following the parameter-based taxonomy defined for Vision-Language Models (VLMs) in Section 2, where

S-VLM	Method	L-VLM: Qwen2VL-7B (Wang et al., 2024) TextVQA	ST-VQA	ChartQA	OKVQA	L-VLM: InternVL2-8B (Chen et al., 2024) TextVQA	ST-VQA	ChartQA	OKVQA	Gains: Max	Average
SmolVLM-500M	ZS	55.3	78.5	56.5	38.2	55.3	78.5	56.5	38.2	3.4	2.4
	MPA	**57.6**$_{(+2.3)}$	**80.3**$_{(+1.8)}$	**59.9**$_{(+3.4)}$	**40.7**$_{(+2.5)}$	**57.7**$_{(+2.4)}$	**80.7**$_{(+2.2)}$	**59.3**$_{(+2.8)}$	**39.9**$_{(+1.7)}$		
TinyLLaVA-2B	ZS	47.1	44.7	12.0	43.6	47.1	44.7	12.0	43.6	15.2	6.8
	MPA	**53.5**$_{(+6.4)}$	**48.7**$_{(+4.0)}$	**24.0**$_{(+12.0)}$	**46.6**$_{(+3.0)}$	**51.9**$_{(+4.8)}$	**49.8**$_{(+5.1)}$	**27.2**$_{(+15.2)}$	**47.2**$_{(+3.6)}$		
InternVL2-2B	ZS	68.0	63.0	63.2	42.7	68.0	63.0	63.2	42.7	5.1	3.0
	MPA	**70.3**$_{(+2.3)}$	**65.5**$_{(+2.5)}$	**68.3**$_{(+5.1)}$	**45.6**$_{(+2.9)}$	**69.5**$_{(+1.5)}$	**65.7**$_{(+2.7)}$	**68.2**$_{(+5.0)}$	**44.6**$_{(+1.9)}$		
InternVL2-4B	ZS	69.1	63.2	73.1	50.5	69.1	63.2	73.1	50.5	4.7	2.1
	MPA	**71.4**$_{(+2.3)}$	**66.6**$_{(+3.4)}$	**73.8**$_{(+0.7)}$	**52.3**$_{(+1.8)}$	**70.3**$_{(+1.2)}$	**67.9**$_{(+4.7)}$	**74.0**$_{(+0.9)}$	**52.0**$_{(+1.5)}$		
Qwen2VL-2B	ZS	70.6	62.5	65.9	47.1	70.7	62.5	65.9	47.1	4.7	2.6
	MPA	**75.1**$_{(+4.5)}$	**67.2**$_{(+4.7)}$	**67.6**$_{(+1.7)}$	**48.9**$_{(+1.8)}$	**72.3**$_{(+1.6)}$	**66.6**$_{(+4.1)}$	**66.9**$_{(+1.0)}$	**48.9**$_{(+1.8)}$		

Table 1: Comparison of our proposed MPA framework performance with the baselines on TextVQA, ST-VQA, ChartQA and OKVQA. The parenthesis (+x) denotes the improvement of +x% over the zero-shot S-VLM by our proposed MPA. The max and average columns show the overall performance gains across all tests for each S-VLM.

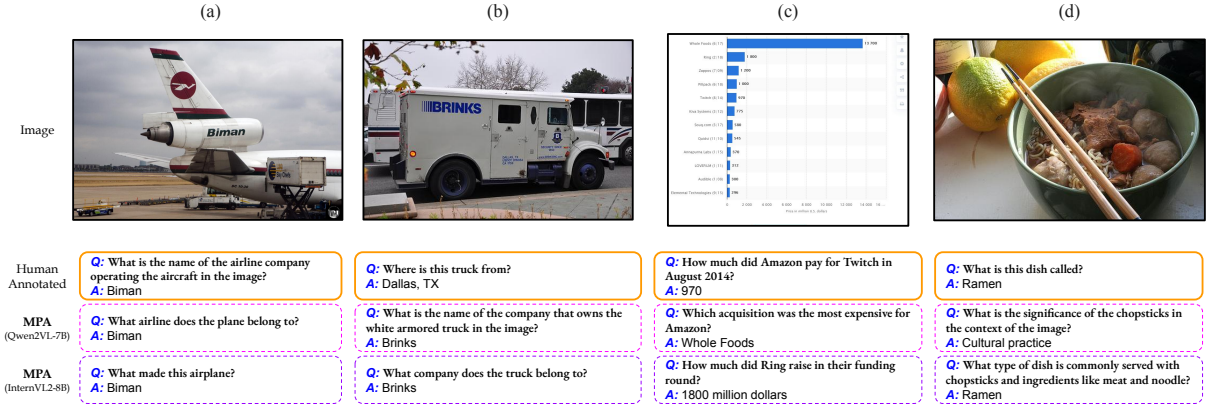


Figure 3: A selection of few pseudo annotations generated by our framework. We further show human annotations from their respective original dataset train splits. **(Best viewed in color)**.

models with $\leq$ 5B parameters are classified as small VLMs (S-VLMs), while those exceeding 5B parameters are large VLMs (L-VLMs) (Lu et al., 2024), we chose five models that range from 500M to 4B parameters as S-VLM, namely SmolVLM-500M (Marafioti, 2024), TinyLLaVA-2B (Zhou et al., 2024), InternVL2-2B (Chen et al., 2024), Qwen2VL-2B (Wang et al., 2024), and InternVL2-4B (Chen et al., 2024); and two open-source models viz. Qwen2VL-7B (Wang et al., 2024) and InternVL2-8B (Chen et al., 2024) and one closed-source model, i.e., GPT-4o (OpenAI, 2024) as L-VLM.

4.1 Results and Discussion

We present the quantitative results of our MPA framework across four datasets evaluated in ten combinations of two L-VLMs and five S-VLMs in Table 1. The results show that MPA consistently improves the performance of all S-VLMs in all datasets with 15.2% maximum and 3.4% average gain in an absolute scale. Here, we analyze the results from the following three key perspectives.

(i) S-VLM family-specific analysis The most noticeable gains are as follows (refer Table 1): TinyLLaVA-2B achieves 27.2% accuracy on ChartQA with our MPA framework, guided by InternVL2-8B, marking an absolute improvement of +15.2% over its original zero-shot performance. Similarly, Qwen2VL-2B, guided by Qwen2VL-7B and InternVL2-4B, guided by InternVL2-8B in our MPA framework achieve +4.7% and +4.7% improvements, respectively, on ST-VQA. On ChartVQA, SmolVLM-500M, guided by Qwen2VL-7B in our MPA framework, improves by +3.4%, while InternVL2-2B, guided by Qwen2VL-7B, gains +5.1%. These results highlight effectiveness of MPA in enhancing the performance of S-VLMs across diverse VQA tasks.

S-VLM	GPT-4o as L-VLM
TinyLLaVA-2B	47.1
TinyLLaVA-2B + MPA	**55.4**$_{(+8.3)}$
Qwen2VL-2B	70.6
Qwen2VL-2B + MPA	**75.4**$_{(+4.8)}$

Table 2: Comparison of MPA-aligned S-VLMs against baseline S-VLMs on TextVQA, with GPT-4o as LVLM.

(ii) VQA Task-specific analysis We observe that TinyLLaVA-2B+MPA aligned with InternVL2-8B

Task	Dataset	Metric	SVLM	SVLM+MPA
OCR	ICDAR2015 (Karatzas et al., 2015)	WRR	31.9	**36.4**(↑ 4.5)
TC	TextCaps (Sidorov et al., 2020)	BLEU-1	7.9	**15.3**(↑ 7.4)
		ROUGE-L	17.4	**20.6**(↑ 3.2)
		CIDEr	8.7	**38.1**(↑ 29.4)

Table 3: **MPA transfers the fundamental capabilities beyond VQA.** In our MPA framework, we use S-VLM: TinyLLaVA-2B, L-VLM: Qwen2VL-7B. Here, **OCR:** visual-text recognition, **TC:** text-aware image captioning. **WRR:** word recognition rate.

achieves a notable +15.2% gain on ChartQA, highlighting our MPA's strength as a knowledge alignment module. In this scenario, it effectively identifies and bridges the knowledge gap between L-VLM and S-VLM for '*complex visual reasoning that involves interpreting charts and graphs*. Improvements on TextVQA (+6.4%) and ST-VQA (+5.1%) further demonstrate MPA's ability to transfer '*visual text understanding*' from larger to smaller models. The modest gain on OKVQA reflects its reliance on external knowledge, which S-VLM inherently lack. While MPA enhances internal knowledge utilization, it cannot fully address such gaps without RAG or fine-tuning on knowledge-rich data. The results validate the effectiveness of MPA within its scope, while highlighting the challenges of knowledge-intensive visual question answering.

(iii) Model size-specific Analysis: MPA improves performance on all model scales, from SmolVLM-500M to InternVL2-4B, demonstrating its versatility. In particular, TinyLLaVA-2B achieves the highest average gain of +6.8 across all tasks, whereas InternVL2-4B shows a comparatively modest improvement of +2.1. We attribute this contrast to two factors: (i) Pretraining data gaps: smaller models like TinyLLaVA-2B benefit more from MPA as it effectively fills missing capabilities through targeted alignment; (ii) Diminishing returns with scale: it is inherently harder to align larger models (4B in this case) that already possess stronger capabilities, in line with scaling laws.

(iv) L-VLM-Specific Analysis: We analyze the effectiveness of different guiding L-VLMs within MPA by computing average gains across five S-VLMs and four VQA datasets. Qwen2VL-7B achieves the highest average improvement of +3.5 points, followed closely by InternVL2-8B with +3.2 points. This suggests that while both models are effective guides, Qwen2VL-7B offers a slightly stronger alignment signal, potentially due to differences in their pretraining objectives or representations. These results highlight that MPA is robust to the choice of L-VLM, yet benefits from stronger or more task-aligned guides.

L-VLM	Status	A ↑	AC ↑	TR ↑	HLS ↑
Qwen2VL-7B	Pre-PI	0.76	0.68	0.8	58
	Post-PI	**0.92**	**0.84**	**0.92**	**74**
InternVL2-8B	Pre-PI	0.74	0.65	0.78	56
	Post-PI	**0.87**	**0.78**	**0.88**	**73**

Table 4: User study on the pseudo-annotations quality: Pre-PI and Post-PI in MPA. **A**: answerability, **AC**: answer correctness, **TR**: task relevancy, **HLS**: Human Likeness Score. Refer Section 4.1.1 for more details.

4.1.1 Ablations and Analysis

We conduct the following ablations and analysis:

(i) How effective is MPA in aligning S-VLMs with closed-source models?: MPA can also leverage powerful closed-source L-VLMs to improve S-VLMs. To assess this, we performed experiments using GPT-4o (OpenAI, 2024) as the guiding L-VLM. As shown in Table 2, MPA consistently improves performance across all aligned S-VLMs, despite having no access to the guiding model's logits or weights. This demonstrates MPA's unique advantage over standard distillation methods, which require full model access. With the expected rise in powerful closed-source models (OpenAI, 2024; Team et al., 2023), such alignment strategies become increasingly valuable. In fact, our results show that integrating powerful L-VLM, e.g. GPT-4o through MPA brings S-VLMs closer or even better in performance to significantly larger models, e.g., MPA-aligned Qwen2VL-2B (75.4%) outperforms Qwen2VL-7B (74.7%).

(ii) Does MPA transfers the fundamental capabilities beyond VQA?: MPA is designed to enhance the VQA performance of S-VLMs by aligning them with L-VLMs, and our results confirm its effectiveness. To examine whether MPA also transfers broader fundamental capabilities such as visual text understanding, we evaluate zero-shot TinyLLaVA-2B and its MPA-aligned counterpart on two different tasks: visual text recognition on ICDAR 2015 (Karatzas et al., 2015) and text-aware image captioning on TextCaps (Sidorov et al., 2020), using Qwen2VL-7B as the guiding L-VLM in MPA. As shown in Table 3, the MPA-aligned model improved text recognition accuracy by 4.5% on an absolute scale and yields notable improvements in captioning metrics such as ROUGE-L and CIDEr. These results suggest that MPA transfers fundamental text understanding capabilities from L-VLMs to S-VLMs beyond the VQA.

(iii) How effective is the role of PI in pseudo-

Method	TextVQA	ST-VQA	ChartQA	OKVQA
LoRA SFT	71.9	63.4	66.1	47.9
Full SFT	71.8	61.7	65.7	47.7
MPA	**75.1**	**67.2**	**67.6**	**48.9**

Table 5: Comparison of few-shot methods vs MPA-aligned Qwen2VL-2B with Qwen2VL-7B as L-VLM. Please note that MPA operates without any human-labeled samples, whereas the other two baselines each use 100 human-labeled samples.

annotation quality correction?: Incorrect annotations may cause models to learn spurious patterns, exhibit biased behavior, and suffer from degraded performance and reliability in downstream tasks. To assess the impact of the PI module on genereted annotation quality, we conducted a user study in which three annotators evaluated 500 randomly sampled pseudo-annotations prior and post PI processing. The evaluation used the following metrics: (a) *Answerability (A)*: 1 if the question is answerable from the image, 0 otherwise; (b) *Answer Correctness (AC)*: 1 if the answer is correct, assuming the question is valid; (c) *Task Relevance (TR)*: 1 if the question aligns with the task, 0 otherwise; and (d) *Human-Likeness Score (HLS)*: percentage of PI-sampled annotations mistaken for human-annotated ones in a mixed set. As shown in Table 4, post-PI annotations exhibited higher quality across all metrics, with more being identified as human-annotated. Figure 3 provides visual evidence by illustrating the high correlation between MPA-generated annotations and human annotated samples. These results validate that PI effectively filters noise and corrects errors, enhancing the overall reliability of MPA-generated annotations.

(iv) How does MPA compare to few-shot supervised baselines? While MPA is designed for a setting where human-labeled traning data is unavailable, obtaining a small labeled set (e.g., 100 samples) is often feasible. In such scenarios, commonly adopted few-shot supervised methods like LoRA-based SFT and full SFT can be applied directly to the S-VLM. To benchmark MPA against these methods, we fine-tune Qwen2VL-2B using both approaches and compare them with MPA-aligned Qwen2VL-2B (using Qwen2VL-7B as L-VLM). As shown in Table 5, MPA consistently outperforms both baselines without labeled supervision, demonstrating high-quality label generation and effective knowledge transfer.

(v) Does PI filtering improve over raw pseudo-labels or full human-labeled data? While our primary focus is on label-free training using MPA, we further investigate the quality of supervision introduced by PI filtering. Specifically, we compare three settings for training Qwen2VL-2B: (i) full human-labeled data (HL), (ii) pseudo-labeled data (PL) from MPA without PI filtering, and (iii) high-quality subset selected by PI that targets the knowledge gap. As shown in Table 6, the PI-selected subset achieves the highest accuracy across all tasks–TextVQA (75.1%), ST-VQA (67.2%), and ChartQA (67.6%), despite using far fewer samples. Interestingly, the performance gain from full human-labeled data over zero-shot baselines is relatively limited. Prior work (Khan et al., 2023) suggests that excessive labeled data can introduce redundancy or noise, reducing the marginal benefit of supervision. This highlights the value of PI filtering in identifying high-utility samples that yield more efficient and effective learning.

Data	Labels	TextVQA		ST-VQA		ChartQA	
		#Samples ↓	Acc. ↑	#Samples ↓	Acc. ↑	#Samples ↓	Acc. ↑
Original	HL	35K	72.7	22K	65.5	28K	66.9
MPA (w/o PI)	PL	21K	73.6	15K	65.8	19K	67.4
MPA	PL	**2K**	**75.1**	1.5K	**67.2**	1.6K	**67.6**

Table 6: Ablation result of using samples from MPA v/s MPA without PI filtering, with Qwen2VL-7B as L-VLM and Qwen2VL-2B as S-VLM inside MPA. HL: Human Labeled. PL: Pseudo Labeled.

Model	Method	**Acc. (%)**
TinyLLaVA-2B	ZS	51.2
TinyLLaVA-2B	MPA	**53.6**$_{(+2.4)}$

Table 7: Performance on Medical VQA (PathVQA). MPA-aligned TinyLLaVA-2B (with Qwen2VL-7B as L-VLM) shows improved cross-domain generalization.

(vi) Beyond standard VQA applicability (Medical VQA): To evaluate MPA's utility beyond standard VQA tasks, we assess its performance in the medical domain using the PathVQA dataset (He et al., 2020). We compare zero-shot TinyLLaVA-2B with its MPA-aligned counterpart, guided by Qwen2VL-7B. We focus on the binary (yes/no) subset of PathVQA, as the open-ended questions often contain highly specialized medical terminology that poses challenges even for large models and may not reflect generalizable reasoning capabilities. As shown in Table 7, MPA yields a gain of +2.4%, demonstrating effective knowledge transfer even in diverse domain-specific settings. These results highlight MPA's ability to generalize across domains without requiring task-specific data or fine-tuning.

(vii) Knowledge Gap Analysis. To better character-

ize the nature of the "knowledge gaps" between S-VLMs and L-VLMs, we manually inspected 100 randomly selected $\mathcal{D}_{PI}$ samples per task. In particular, we compared Qwen2VL-2B and Qwen2VL-7B within the MPA framework. We categorize the dataset-wise knowledge gaps into key categories, which are summarized in Table 8. Note that *Correct but Verbose* and *Noisy/Task-Irrelevant* cases are excluded from the knowledge-gap categories, as they do not represent fundamental reasoning shortcomings.

Further, we provide representative examples that illustrate these knowledge-gap categories:

1. **TextVQA/ST-VQA:** (i) *Shallow OCR grounding* (Fig. 5 (a)): "What word is printed under interior design on the book in the middle?" — S-VLM outputs "para" from a nearby visible region instead of grounding to the queried location. (ii) *Noisy or hallucinated OCR* (Fig. 7 (d)): "What company's logo is in the black box in the upper left?" — S-VLM hallucinates "Burberry" without actually reading the text.

2. **ChartQA:** (i) *Entity misalignment* (Fig. 8 (d)): "Who was the leading goal scorer for Celtic FC as of September 2020?" — the S-VLM retrieves an incorrect player name that is not aligned with the queried entity. (ii) *Conditional/chart understanding error* (Fig. 5 (c)): "Which year yielded the smallest difference between men and women students?" — S-VLM fails to detect the year with the minimum gap between the trend lines. (iii) *Trend misinterpretation* (Fig. 8 (b)): "Does the life expectancy decrease over the years?" — S-VLM misinterprets the slope changes.

3. **OKVQA** (i) *Lack of internal knowledge grounding* (Fig. 9 (b)): "At what speed does this animal run?" — S-VLM fails to answer, while the MPA-aligned model succeeds without external knowledge, highlighting the shallow grounding of the S-VLM. (ii) *Visual guesswork* (Fig. 9 (c)): "What is the name of the floor pattern?" — S-VLM guesses "diamond" from vague cues instead of leveraging the consistent checkered pattern.

5 Conclusion and Future Work

In this work, we introduced the Model Parity Aligner (MPA), a novel framework that enhances small vision-language models (S-VLMs) by leveraging unlabeled images and effective knowledge transfer from large vision-language models (L-VLMs). Unlike traditional knowledge distillation techniques that rely on labeled data and access to large model logits, MPA employs pseudo-labeling with quality assessment, ensuring that small models learn from high-confidence supervision while avoiding error propagation. Our experiments across four diverse VQA benchmarks, viz. TextVQA, ST-VQA, ChartQA and OKVQA demonstrate that MPA consistently improves S-VLM performance, making them more viable for real-world applications with limited resources.

Despite these improvements, there still remains a gap between S-VLMs and L-VLMs that highlights the need for further advancements. As future work, we aim to explore more robust knowledge alignment strategies, including iterative refinement of pseudo-labels, leveraging diverse sources of unlabeled data, and integrating multi-step reasoning from L-VLMs into S-VLMs training. Additionally, extending MPA to tasks beyond visual question answering could further enhance its applicability. We view MPA as a first step toward achieving model parity in vision and language models via targeted knowledge alignment, and firmly believe that it shall open up future research avenues for more efficient and capable small models for vision-language tasks.

Dataset	Error Category	# Samples
TextVQA/ST-VQA	Shallow OCR grounding	33
	Noisy or hallucinated OCR	53
	Correct but Verbose	5
	Noisy/Task-Irrelevant samples	9
ChartQA	Entity misalignment	55
	Conditional/chart understanding errors	17
	Trend misinterpretation	14
	Correct but Verbose	8
	Noisy/Task-Irrelevant samples	6
OKVQA	Lack of internal knowledge grounding	23
	Visual guesswork	58
	Correct but Verbose	4
	Noisy/Task-Irrelevant samples	15

Table 8: Distribution of error categories across datasets in our manual inspection of 100 $\mathcal{D}_{PI}$ samples per task. Note that *Correct but Verbose* and *Noisy/Task-Irrelevant* are not true knowledge-gap categories.

Limitations

Our proposed MPA framework depends on access to a large vision-language model (L-VLM) for generating and validating pseudo-annotations. In even stricter resource-constrained settings, this may limit applicability of MPA. Further, when leveraging proprietary closed-source models via commercial APIs, reproducibility and transparency may

be compromised due to limited insight into model behavior and potential changes in API responses over time. Our experiments also focus primarily on English-language datasets and VQA-related tasks; generalization to multilingual, or more complex reasoning tasks remains an open direction.

Ethical Considerations and Broader Impact

In this work, we used open-source datasets which may contain social or cultural biases. The proposed framework also depends on outputs from large-scale vision-language models (L-VLMs), which are known to occasionally generate hallucinated or biased content. Although the Parity Identifier (PI) module is designed to filter out low-quality or incorrect annotations, it cannot entirely eliminate inherited biases from the underlying L-VLM. Further, this work involves a human evaluation study in which three annotators were employed to assess the quality of pseudo-annotations generated by our MPA framework. All annotators were compensated fairly in accordance with local wage norms. They were not exposed to harmful, offensive, or sensitive content, and no personally identifiable information was collected at any stage of the study.

Broader Impact: The proposed MPA framework enables efficient training of small vision-language models (S-VLMs) using only unlabeled data, reducing reliance on expensive human annotations. By transferring capabilities from large vision-language models (L-VLMs) to compact models, MPA makes high-performing multimodal systems more accessible in low-resource settings. This democratization of vision-language technology can benefit real-world applications in healthcare, agriculture, and accessibility, particularly in regions with limited compute or labeled data. Furthermore, the proposed approach encourages the development of scalable alignment strategies that can generalize to diverse, resource-constrained communities.

Acknowledgments

This work was partly supported by the National Language Translation Mission (NLTM): Bhashini project by the MeitY, Government of India. Abhirama Subramanyam Penamakuri was supported by the PMRF fellowship, MoE, Government of India.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. Weaqa: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *ICCV*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. 2024. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*.

Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *European conference on computer vision*, pages 95–112. Springer.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE international conference on computer vision*, pages 1409–1416.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi.

2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

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

HuggingFaceTB. 2023. smolllm: A collection of small language models. https://huggingface.co/collections/HuggingFaceTB/smollm-6695016cad7167254ce15966. Accessed: 2025-03-06.

Soumya Jahagirdar, Shankar Gangisetty, and Anand Mishra. 2021. Look, read and ask: Learning to ask questions by reading text in images. In *International Conference on Document Analysis and Recognition*, pages 335–349.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6485–6494.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th international conference on natural language generation*, pages 198–202.

Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. 2021. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukás Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 1156–1160.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. 2023. Q: How to specialize large vision-language models to data-scarce vqa tasks? a: Self-train on unlabeled images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15005–15015.

Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. 2021. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6361.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 24872–24895.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2008–2018.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

Andres Marafioti. 2024. Smolvlm - small yet mighty vision language model.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *ACL*.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *NeurIPS*.

Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. 2018. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. 2024. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, and 1 others. 2024. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *CVPR*.

Victoria Stodden. 2020. The data science life cycle: a disciplined approach to advancing data science as a science. *Commun. ACM*, 63(7):58–66.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. 2024. Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. *arXiv preprint arXiv:2402.04616*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. 2022. Guiding visual question generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1640–1654.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847.

Guo-Hua Wang, Yifan Ge, and Jianxin Wu. 2021. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195.

Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F. JáJá, and Larry Davis. 2022. TAG: boosting text-vqa via text-aware visual question-answer generation. In *BMVC*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. 2020. Feature normalized knowledge distillation for image classification. In *European conference on computer vision*, pages 664–680. Springer.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. 2024. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. Automatic generation of grounded visual questions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4235–4243.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Additional Analysis

(i) Additional comparisons: utility of PI filtering over raw pseudo-labels. We extend the analysis of PI filtering by reporting the results of MPA (w/o PI) across all S-VLMs with Qwen2VL-7B as the guiding L-VLM within MPA. As shown in Table 9, MPA consistently outperforms MPA (w/o PI) across all tasks and S-VLMs, despite using far fewer training samples (for instance, ∼2K vs. ∼21K for TextVQA). These results reinforce the utility of PI filtering in isolating knowledge-gap samples that provide more efficient and targeted supervision.

(ii) Expanded comparison on OCR and text-aware captioning tasks. In Table 3, we examined whether MPA-trained models can improve fundamental capabilities such as OCR and text-aware image captioning, even without direct supervision. We further evaluate this setting by comparing against models fine-tuned on the original human-labeled training splits of TextVQA; the results are presented in Table 10. As shown, MPA not only improves over the zero-shot baseline but also surpasses models trained with human-labeled annotations. This highlights that the gains stem from the effectiveness of the MPA pipeline, rather than from overlap between benchmarks, and demonstrates that MPA successfully transfers core visual-linguistic capabilities in a label-free manner.

(iv) Computational and API cost of PA and PI: MPA is a one-time pipeline where each image is processed by the L-VLM during the PA phase, and each generated (image, question) pair is passed once through the L-VLM and S-VLM during the PI phase. For open-source L-VLMs like Qwen2VL-7B deployed locally, this is computationally lightweight: on a machine with 3 A6000 (48GB) GPUs, generating approximately 21K pseudo-annotations (e.g., for TextVQA) takes around 4-6 hours end-to-end. Further, the PI step takes another 2-3 hours to identify the samples that represent the knowledge gaps. Alternatively, while using GPT-4o via API, we estimate the total cost of PA + PI for a single S-VLM–task pair to be around $11, making MPA a highly cost-effective label-free alternative to supervised training.

B Dataset Details

TextVQA consists of 28K images with 45K manually annotated question-answer pairs. It is split into 21K images with 35K questions for training, 3K images with 3.7K questions for validation, and a private test set. Since the testset is private, for this dataset, we report all the result on validation set. ST-VQA contains 23K images and 31K questions, with 16K images and 22K questions for training, and 2.8K images with 4K questions for testing. ChartQA includes 21.6K charts with 32.3K question-answer pairs, split into 19K charts with 28K questions for training, 1K charts with 1.8K questions for validation, and 1.6K charts with 2.5K questions for testing. OKVQA consists of 14K images with 14K questions, divided into 9K questions for training, 5K for testing.

S-VLM	Samples	TextVQA	ST-VQA	ChartQA	OKVQA
TinyLLaVA-2B	MPA (w/o PI)	56.4	79.1	57.5	39.2
	MPA	**57.6**	**80.3**	**59.9**	**40.7**
TinyLLaVA-2B	MPA (w/o PI)	52.1	46.3	23.3	44.6
	MPA	**53.5**	**48.7**	**24.0**	**46.6**
InternVL2-2B	MPA (w/o PI)	69.0	64.5	66.7	44.0
	MPA	**70.3**	**65.5**	**68.3**	**45.6**
InternVL2-4B	MPA (w/o PI)	69.8	65.1	72.9	51.2
	MPA	**71.4**	**66.1**	**73.8**	**52.3**
Qwen2VL-2B	MPA (w/o PI)	73.6	65.8	67.4	47.2
	MPA	**75.1**	**67.2**	**67.6**	**48.9**

Table 9: Additional results for MPA vs. MPA (w/o PI) across all S-VLMs, using Qwen2VL-7B as the L-VLM inside MPA.

C Implementation Details

We implement our method using PyTorch. Majority of the chosen S-VLMs and L-VLMs employed in our propsed method MPA, we use their original code-base repositories and/or their Huggingface implementations depending on the ease of reproducibility. Parity leveler (Section 3.3) module trains the entire S-VLM on the samples obtained from the PI module (Section 3.2) for one epoch, for all the benchmark datasets. Hyperparameters used by the PL module for different S-VLMs are summarized in Table 11. All our experiments are conducted on a machine with three Nvidia A6000 GPUs (48 GB each). For every L-VLM and S-VLM combination, it took approximately, 5-12 GPU hours for entire MPA, for one dataset. We use `gpt-4o-2024-11-20` (OpenAI, 2024) for our closed-source L-VLM ablation.

D Prompts used

In this section, we provide the VLM prompts used in the PA module (Section 3.1) to generate pseudo-annotations for all four datasets:

Task	Dataset	Metric	S-VLM (Zero-shot)	S-VLM (HL)	S-VLM (MPA)
OCR	ICDAR2015	WRR	31.9	33.2	**36.4** (↑ 4.5)
TC	TextCaps	BLEU-1	7.9	13.4	**15.3** (↑ 7.4)
		ROUGE-L	17.4	18.3	**20.6** (↑ 3.2)
		CIDEr	8.7	34.6	**38.1** (↑ 29.4)

Table 10: Comparison of OCR and text-aware captioning performance. Despite using no ground-truth labels, MPA outperforms both the zero-shot baseline and models trained on human-labeled data (HL).

S-VLM	Batch Size	LR
Qwen2VL-2B	16	1e-5
InternVL2-2B	16	4e-5
InternVL2-4B	6	4e-5
SmolVLM-500M	16	1e-4
TinyLLaVA-2B	16	1e-4

Table 11: Hyperparameters used in the parity leveler module (Section 3.3) for each S-VLM.

PA prompt for: TextVQA and ST-VQA

<image(**I**)>
The objective is to generate a question-answer pair for a Textual Visual Question Answering (Text-VQA) task. Your task is to create a contextually relevant question that directly relates to the image's content, incorporating reasoning or direct references to the text, and its correct answer.
Output:
- Question: A natural language question grounded in the image's content and text.
- Answer: A concise response (single word, phrase, or Yes/No) derived from the text or reasoning based on it.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

PA prompt for: ChartQA

<chart image (**I**)>
The objective is to generate a question-answer pair for a Chart Visual Question Answering (ChartVQA) task. Your task is to create a contextually relevant question that directly relates to the content of a given chart, incorporating reasoning based on the visualized data.
Output Requirements:
- Question: A natural language question grounded in the chart's content, requiring numerical reasoning, trend analysis, or data lookup.
- Answer: A concise response (single word, number, phrase, or Yes/No) derived from the chart's data.
Guidelines for Question Generation:
1. Direct Lookup Questions – extracting specific values from the chart.
2. Comparison Questions – comparing values between different categories.
3. Trend & Pattern Recognition – identifying increases, decreases, or correlations in the data.
4. Inference-Based Questions – requiring reasoning beyond direct value lookup.
Ensure the question is meaningful and the answer is accurate based on the chart data.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

PA prompt for: OKVQA

<image(**I**)>
The objective is to generate a question-answer pair for a Knowledge-based Visual Question Answering (K-VQA) task. Your task is to create a contextually relevant question that directly relates to the image's content while requiring external world knowledge to answer correctly, and its correct answer.
Output Requirements:
- Question: A natural language question grounded in the image's content but requiring reasoning beyond direct perception, incorporating real-world knowledge.
- Answer: A single-word response based on general world knowledge.
Guidelines for Question Generation:
1. Object & Scene Understanding – identifying objects or actions in the image and connecting them to broader knowledge.
2. Commonsense Reasoning – requiring logical deductions about the scene.
3. Cultural & Historical Context – related to well-known historical events, traditions, or cultural references.
4. Scientific & Factual Knowledge – involving basic physics, biology, geography, or general knowledge.
5. Everyday Life & Social Understanding – questions about daily activities, professions, or human behaviors.
Ensure that the generated question is meaningful and requires external knowledge beyond just the image's visual content.
Assistant: Question: $\tilde{\mathbf{Q}}$, Answer: $\tilde{\mathbf{A}}$

Note that, to ensure fair comparison, the pseudo-annotation prompts are same for all variants of L-VLMs used. Further, the prompt we used for QA is '`Answer the following question in a single word or phrase`', which is common for all datasets across all S-VLMs.

E Qualitative Results

Figure 5 presents a selection of examples where MPA alignment enables S-VLM to correct errors made by the original zero-shot S-VLM. From a rigorous examination of the results, we find that MPA significantly improves performance in visual text reasoning, plot interpretation, and knowledge-based question answering. Further, we show additional qualitative samples for showing zero-shot SVLM versus MPA-aligned S-VLM across all four

Figure 4: **Left two examples:** Pseudo-annotations discarded by PI module as they do not constitute knowledge-gap. **Right two examples:** Pseudo-annotations discarded by PI module as they are noisy annotations.

datasets: TextVQA in Figure 6, ST-VQA in Figure 7, ChartQA in Figure 8 and OKVQA in Figure 9.

Furthermore, in Figure 4, we show selected examples that do not represent a disparity between S-VLM and L-VLM ((a), (b)), and another set of examples that are noisy annotations ((c), (d)), both of which are discarded by the PI module.

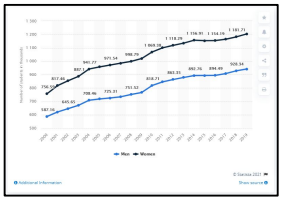
	(a)	(b)	(c)	(d)
Image				
Q	What word is printed under interior design on the book in the middle?	Where was this product made?	Which year yielded the smallest difference between men and women students?	What effect on the ocean does this platenary body cause?
SVLM	Para	USA	2005	Moon
SVLM+MPA (Ours)	**Inspirations**	**UK**	**2000**	**Tide**

Figure 5: A selection of results showing zero-shot SVLM versus MPA-aligned SVLM. MPA config: S-VLM: Qwen2VL-2B, L-VLM: Qwen2VL-7B. Green and red text correspond to correct and incorrect answers, respectively. **(Best viewed in color)**

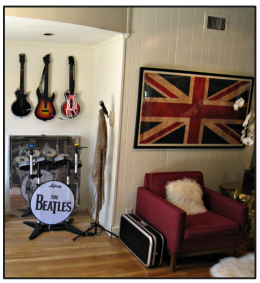
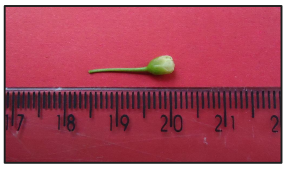
	(a)	(b)	(c)	(d)
Image				
Q	What brand is depicted above the beatles logo?	Which way to turn to porchester terrace?	What is the first even number?	What number and letter are found under the green icon?
SVLM	Gibson	Left	17	6
SVLM+MPA (Ours)	**Ludwig**	**Right**	**18**	**F6**

Figure 6: Few more results from TextVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM.

	(a)	(b)	(c)	(d)
Image				
Q	What is the punishment for honking?	What letters come after the letters ATV/ on the same button?	What is printed on the right side of the clock?	What company's logo is in the black box in the upper left?
SVLM	$150 penalty	DVD	Letters	Burberry
SVLM+MPA (Ours)	**$350 penalty**	**DTV**	**1240 KC**	**Gucci**

Figure 7: Few more results from STVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. **(Best viewed in color)**

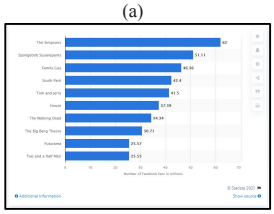
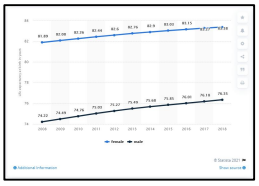
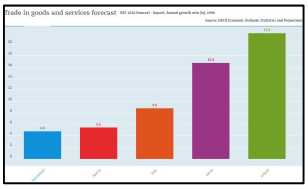
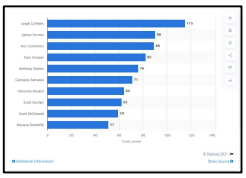
	(a)	(b)	(c)	(d)
Image				
Q	How many facebook fans did the simpsons have as of June 2011?	Does the life expectancy decrease over the years?	How many places are mentioned in the graph?	Who was the leading goal scorer for Celtic FC as of September 2020?
SVLM	26.1 million	Yes	10	Scott Sinclair
SVLM+MPA (Ours)	**62 million**	**No**	**5**	**Leigh Griffiths**

Figure 8: Few more results from ChartQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. **(Best viewed in color)**

	(a)	(b)	(c)	(d)
Image				
Q	The chef is holding a pizza in the photo so what type of restaurant does this suggest he may be cooking at?	At what speed does this animal run?	What is the name of the floor pattern?	Can you guess the model of tv shown in this picture?
SVLM	Pizza	10 mph	Diamond	No
SVLM+MPA (Ours)	**Italian**	**30 mph**	**Checkered**	**LG**

Figure 9: Few more results from OKVQA showing the efficiency of MPA-aligned S-VLM over baseline S-VLM. Green and red text correspond to correct and incorrect answers, respectively. **(Best viewed in color)**