arXiv:2409.14215v1 [cs.CV] 21 Sep 2024

@BENCH: Benchmarking Vision-Language Models for Human-centered Assistive Technology

Xin Jiang[1,3,*], Junwei Zheng[1,*], Ruiping Liu[1], Jiahang Li[2], Jiaming Zhang[1,†], Sven Matthiesen[2], Rainer Stiefelhagen[1]

[1]CV:HCI, Karlsruhe Institute of Technology, [2]IPEK, Karlsruhe Institute of Technology, [3]Li Auto Inc.

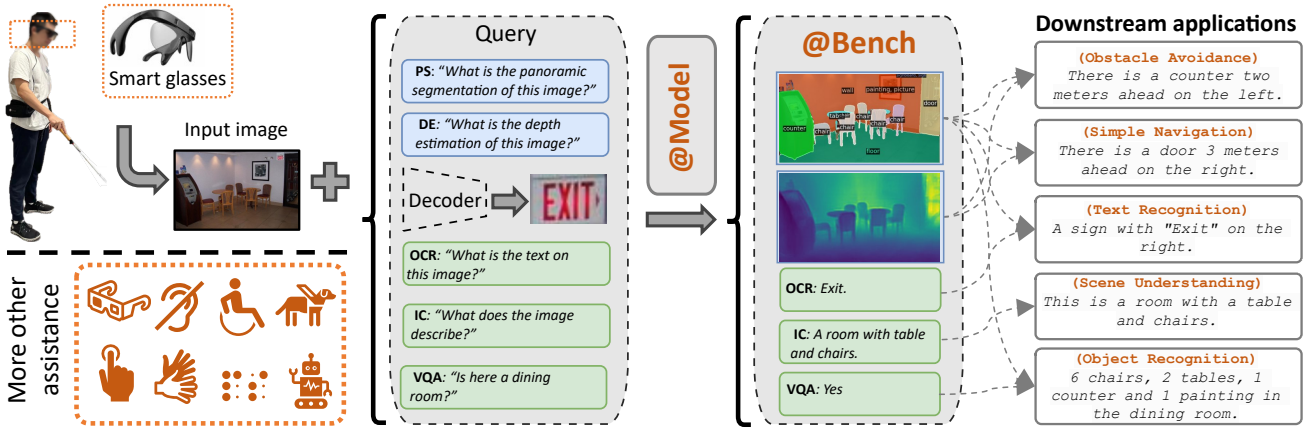

Figure 1. **Overview of our Assistive Technology Model (@MODEL) and Benchmark (@BENCH).** @MODEL can perform vision-language tasks all at once, including: Panoptic Segmentation, Depth Estimation, Image Captioning, Optical Character Recognition and Visual Question Answering. All tasks of @BENCH are selected by People with Visual Impairments (PVIs) to evaluate VLMs for AT.

Abstract

As Vision-Language Models (VLMs) advance, Assistive Technologies (ATs) for helping People with Visual Impairments (PVIs) are evolving into generalists, capable of performing multiple tasks simultaneously. However, benchmarking VLMs for ATs remains under-explored in the literature. To bridge this gap, we first create a novel ***AT benchmark (@BENCH)****. Guided by a pre-design user study with PVIs, our benchmark includes the five most crucial vision-language tasks:* Panoptic Segmentation, Depth Estimation, Optical Character Recognition (OCR), Image Captioning, *and* Visual Question Answering (VQA). *Additionally, we propose a novel* ***AT model (@MODEL)*** *that addresses all tasks simultaneously and can be expanded to more assistive functions for helping PVIs. Our framework exhibits outstanding performance across tasks by integrating multi-modal information, and it offers PVIs a more comprehensive scene understanding. Extensive experiments prove the effectiveness and generalizability of our framework.*

*Equal contribution.
†Corresponding author (e-mail: jiaming.zhang@kit.edu).
All codes will be made publicly available at ATBench.

1. Introduction

Assistive Technologies (ATs) for People with Visual Impairments (PVIs) have witnessed significant advancements in recent years where computer vision and natural language processing play an important role. Previous works [1, 10, 11, 46] focus on providing PVIs with specific and limited functionalities, such as navigation [1], obstacle avoidance [46], image captioning [11], *etc.* However, existing methods face challenges in efficiently processing multiple tasks simultaneously. Specifically, existing approaches struggle to accurately interpret complex scenes, which is essential to meet the needs of PVIs. Additionally, they often provide less contextually relevant information for scene description. Recently, the generalist Vision-Language Models (VLMs) [24, 42, 51] have been proposed and shown great potential in multi-tasking and revolutionizing the next-generation assistive systems. Regarding Vision-Language (VL) tasks, these models benefit from the dividend in both computer vision and natural language processing domains, facilitating collaboration between visual and language tasks for a more comprehensive understanding of the surrounding environment. Nonetheless, benchmarking VLMs for ATs is

still under-explored. The previous Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) benchmarks [23, 48] have limitation in two aspects. Firstly, while the benchmarks focus on language-specific and cross-modal tasks, they seldom address pure vision-specific tasks, which are crucial in the context of visual impairments. Secondly, these benchmarks tend to prioritize general applicability over the specific needs of individuals with visual impairments. Therefore, we raise the research question: ***Are VLMs ready for empowering assistive technology for helping People with Visual Impairments (PVIs)?***

To answer this question, we propose a new VL-based AT benchmark (@BENCH) as a platform for evaluating VLMs for visually impaired assistance. To involve the target group in shaping the benchmark, we conduct a number of questionnaires via a human-in-the-loop process with seven participants who are blind or have low vision, so as to understand the practical demands of PVIs. Based on the feedback of the user study, we introduce five tasks ranked by blind users according to the level of interest, frequency of usage, and level of importance. As presented in Fig. 1, the most assistance-related tasks include: *Panoptic Segmentation (PS)*, *Depth Estimation (DE)*, *Optical Character Recognition (OCR)*, *Image Captioning (IC)*, and *Visual Question Answering (VQA)*. Beyond the range of tasks, performance and efficiency stand as crucial considerations in AT for PVIs. Therefore, we introduce an evaluation framework for assessing the trade-off between efficiency and performance in generalist VLMs.

With this benchmark in place, we propose a novel AT model (@MODEL) that use *task-specific prompt* to combine these 5 uni-modal or cross-modal tasks and realize the paradigm of multi-task training. Thanks to this, @MODEL can use one suit of parameters to implement multiple tasks. It is crucial to significantly reduce the number of parameters, and it will be possible to deploy one model and one suit of weights on the portable device for PVIs.

To summarize, we present the following contributions:

(1) **Human-in-the-loop User Study.** As a part of PVIs-specific design, it is necessary to investigate the needs of the target group. We conduct a participatory user study for the sake of understanding the most related tasks, enabling a user-driven design.

(2) **Vision-Language Benchmark for Assistive Technology.** We release a new benchmark with five representative VL tasks close to the daily life of PVIs. Other complex functionalities can be derived from these tasks, such as obstacle avoidance. We further evaluate the efficiency-performance trade-off on @BENCH.

(3) **Generalist Vision-Language Model for Assistive Technology.** We propose a new end-to-end baseline @MODEL for addressing multiple tasks in @BENCH all at once. The model achieves competitive performance compared with state-of-the-art methods.

(4) **One Suit of Weights for All Tasks.** Benefiting from multi-task training, our model is capable of concurrently executing all tasks with a unified set of weights, resulting in a significant reduction in the number of parameters and computational cost.

2. Related Work

2.1. Assistive Technologies for the Blind

A common goal in ATs is to develop artificial intelligent systems via vision-language algorithms to help PVIs. VizWiz, introduced by Bigham *et al.* in 2010 [3], presents the first multi-task datasets and artificial intelligence challenges originating from PVIs. These include over 10 tasks, such as VQA [10], image captioning [11], object detection [33] and object classification [2], *etc.* It covers various scenes in the daily life of PVIs and provides valuable data for the research of ATs. Other task-specific datasets [37, 47] focus on the recognition of obstacles and tactile paving, further contributing to this field. Based on the PVIs-oriented and general datasets, most work used visual model to address the daily challenges encountered by PVIs. For example, previous works [1, 9, 39, 46] employ visual tasks such as detection, segmentation and depth estimation to accomplish avoidance, navigation, privacy protection, *etc.* Some other works have focused on language-modal or cross-modal tasks, such as OCR [25, 30], image captioning [11], and VQA [10]. Compared to these existing methods, we conduct a human-in-the-loop study to design a unified multi-modal benchmark for evaluating VLMs for ATs.

2.2. Generalist Vision-Language Models

Generalist VLMs have witnessed remarkable advancements, driven by breakthroughs in deep learning techniques [31,42,51]. Developing a generalist model for multiple tasks poses unique challenges due to the heterogeneous inputs and outputs, including RGB images, depth maps, binary masks, bounding boxes, language, etc. Previous methods MetaLM [12] and PaLI [5] use language models as general-purpose interfaces to various foundation models. GLIPv2 [45] unifies both localization and VL understanding tasks as grounded vision-language tasks. OFA [42] and Unified-IO [24] introduce a Seq2Seq framework for the unification of I/O, architectures, tasks, and modalities. X-Decoder [51] can predict pixel-level segmentation and language tokens through a generalized decoding model. Uni-Perceiver v2 [22] formulates different tasks as a unified maximum likelihood estimation problem without any task-specific fine-tuning. However, these methods are either unable to perform multi-task training, or they focus too much on language-modal tasks or cross-modal tasks, while ignor-

ing vision-modal tasks that are important for PVIs, such as segmentation and depth estimation. Therefore, we propose a new method that can comprehensively consider and balance multiple assistance-related uni-modal and cross-modal tasks, and can use one suit of weights for all tasks.

2.3. Benchmarks for Vision-Language Models

To evaluate vision-language systems, Zhou *et al.* [50] propose a multi-task multi-dimension benchmark for Vision Language Pretraining (VLP) models. Su *et al.* [36] introduce the GEM benchmark, a multi-modal benchmark that focuses on both image-language tasks and video-language tasks. Recently, many LVLMs benchmarks [23, 48] have emerged to more comprehensively evaluate the fine-grained capabilities of models. However, a vision-language benchmark for ATs and PVIs is lacking in the literature. To bridge this gap, we introduce @BENCH , a benchmark that includes realistic multi-modal tasks closely relevant to the daily lives of PVIs. Therefore, @BENCH is designed to serve as a fundamental benchmark for evaluating VLMs in the realm of ATs.

3. @BENCH: Assistive Technology Benchmark

@BENCH is a pioneering multi-modal benchmark tailored specifically for the domain of Assistive Technology (AT). The primary target of @BENCH is to establish a comprehensive and standardized evaluation platform for vision-language models in the context of helping PVIs.

In this section, we introduce the detail of the user study (Sec. 3.1), which helps us identify important tasks that are closely related to PVIs. We then provide an overview (Sec. 3.2) of the tasks, datasets and metrics encompassed within the @BENCH framework. Subsequently, we introduce the vital dimension for assessing the performance of VLMs for ATs: efficiency-performance trade-off (Sec. 3.3).

3.1. User-centered Study

Organization. To build the AT benchmark, we conducted a pre-study with 2 accessibility experts to develop a reasonable questionnaire. Based on their suggestions and multiple discussions, we further organized a user-centered study with 7 participants who are blind or have low vision. The goal was to identify which vision-language tasks are beneficial for the target group and meet their requirements.

Questionnaire. To enhance the rationale of our benchmark design, we conducted a questionnaire session with the participants. In this questionnaire, we presented 8 functions relevant to the daily lives of PVIs in Table 1: (1) *obstacle avoidance*, (2) *indoor distance estimation*, (3) *object recognition*, (4) *object location*, (5) *text recognition*, (6) *surroundings understanding*, (7) *scene recognition* and (8) *visual Q&A for surroundings*. At the begining of the user study, we first explained the specific concepts of different functions and related usage scenarios to the participants. Then, they were asked to rate each function based on 3 criteria: (1) level of interest, (2) frequency of usage, and (3) importance in their daily life from 1 (lowest) to 5 (highest). In addition, each function has and can be performed by using a corresponding uni-modal or cross-modal task.

Quantitative Result. Each participant rated each function from 1 to 5 based on these three criteria. After collecting the scores of each participant, we averaged the scores of the 7 participants to represent the score of the function, and finally used the total score of the 3 criteria as the score of each function. And we use the total score as a basis to select relevant functions and tasks. The ratings were then aggregated and summarized, as illustrated in Table 1. Functions 5 and 3 have the highest scores. PVIs think text recognition and object recognition are the most important in their daily life. Function 7 has the lowest score and has a large gap with other functions. At the same time, functions 1, 2, 4, 6, and 8 have similar scores. Therefore, we did not consider the tasks corresponding to function 7 in the benchmark, and retained the tasks related to all the remaining functions. According to the study, the selected functions are: panoptic segmentation, depth estimation, OCR, image captioning and VQA.

3.2. Assistive Tasks

Guided by the user study, @BENCH contains 5 tasks that are extremely relevant to the daily lives of PVIs. We give an overview of all tasks and the corresponding datasets in Table 2, and describe the details as follows:

Panoptic Segmentation. It is the task that combines both semantic segmentation and instance segmentation, aiming to simultaneously recognize all object instances in an image and segment them by category, helping blind people perceive the surroundings more accurately. We opt to ADE20K [49], which contains more than 27K images spanning 365 different scenes with totally 150 semantic categories, including indoor, outdoor, urban, rural, *etc.*, basically covering almost all daily life scenes and common objects of PVIs. And we use Panoptic Quality (PQ) [20] to measure the performance.

Depth Estimation. It is the task of measuring the distance of each pixel relative to the user's camera. According to user study, the majority of PVIs spend most of their time indoors. Therefore, we choose NYU v2 [27], which includes indoor scenes. We evaluate with the RMSE metric.

Optical Character Recognition. It is the conversion of images of typed, handwritten or printed text into machine-encoded text. We know text recognition is the most important function for the PVIs from user study, so we select two widely-used, large synthetic datasets for training: MJSynth (MJ) [14] and SynthText (ST) [14]. We evaluate recognition accuracy on 6 datasets, covering various text scenar-

	Function	Related Task	Level of Interest 1 (low)– 5 (high)	Frequency of Usage 1 (low) – 5 (high)	Importance 1 (low) – 5 (high)	Total
1	Obstacle Avoidance	Panoptic Segmentation	3.00	2.71	2.43	08.14
2	Indoor Distance Estimation	Depth Estimation	3.14	2.43	2.43	08.00
3	Object Recognition	Panoptic Segmentation	3.86	3.71	4.00	11.57
4	Object Location	Panoptic Segmentation	3.29	3.14	3.29	09.72
5	Text Recognition	OCR	4.57	4.43	4.43	13.43
6	Surroundings Understanding	Image Captioning	3.43	2.86	2.71	09.00
7	Scene Recognition	Scene Recognition	2.14	1.71	1.86	05.71
8	Visual Q&A	Visual Question Answering	3.57	3.43	3.43	10.43

Table 1. **Quantitative result of the user study.** Potential functions can be achieved via related tasks, which are listed in the questionnaires for the user study. Note: all scores are averages across 7 participants.

Task	Dataset	#Train	#Val	#Test	Metric
PS	ADE20K	25,574	2,000	2,000	PQ
DE	NYU v2	24,230	654	654	RMSE
OCR	MJ, ST, 6 OCR	15,895,356	7,507	7,507	Accuracy
IC	VizWiz_Cap	23,431	7,750	8,000	BLEU-1/CIDEr
VQA	VizWiz_VQA	20,523	4,319	8,000	Accuracy
$\sum$	@BENCH	15,989,114	22230	26161	

Table 2. **Statistic of pre-selected tasks and datasets in @BENCH**. Note that some datasets do not have a *test* subset, so use *val* subset for evaluation. The 6 OCR datasets are IC13, IC15, IIIT5K, SVT, SVTP and CUTE.

ios: ICDAR 2013 (IC13) [17], ICDAR 2015 (IC15) [16], IIIT5K-Words (IIIT5K) [26], Street View Text (SVT) [41], Street View Text-Perspective (SVTP) [29], and CUTE80 (CUTE) [34].

Image Captioning. It is a challenging task that involves generating human-like and coherent natural language descriptions for images. The task is to comprehend visual content and express in natural language that is descriptive and contextually relevant, which allows PVIs to have an overall understanding of their surroundings. To better align with the daily experiences of PVIs, we opted for the VizWiz_Cap [11], collected from the perspective of PVIs. We use BLEU-1 [28] and CIDEr [38] as evaluation metrics.

Visual Question Answering. It requires the model to take as input an image and a free-form, open-ended, natural language question. It produces or selects a natural language answer as output [50]. In our work, we found that PVIs expressed significant interest in this task. By posing questions, they can experience and comprehend the unseen world around them, providing them with a novel and enriching experience. We select VizWiz_VQA [10] and use the publicly released VizWiz_VQA evaluation scripts in @BENCH.

3.3. Efficiency-Performance Trade-off

In designing models for assistive systems, an optimal balance between efficiency and performance is essential. Normally, the performance of VLMs can be easily measured by task-specific metrics. For efficiency, there are few common choices: the number of parameters, FLOPs and inference time. To compare with previous methods, we evaluate efficiency through the number of parameters.

In sum, @BENCH is designed specifically for multi-modal, multi-task scenarios and tailored to assistive systems for PVIs. All tasks are closely tied to the needs of PVIs community. @BENCH not only prioritizes performance but also places emphasis on efficiency-performance trade-off. We aspire for this benchmark to serve as a cornerstone for researchers within PVIs assistance community, encouraging exploration of multi-modal models' applications in ATs.

4. @MODEL: Assistive Technology Model

Based on X-Decoder, we propose @MODEL, the first generalist model to support all these assistance-related vision-language tasks. As show in Fig. 2, the overall model is built on top of a image encoder for extracting image features, two text encoders that share parameters for extracting text features, and a transformer decoder with generic latent queries and textual queries. @MODEL has two types of output: (1) pixel-level output for dense prediction, such as panoptic segmentation and dense estimation and (2) token-level output for a diverse set of language-related vision tasks, such VQA, image captioning and OCR.

Unified Multi-task Architecture. X-Decoder [51] includes only two tasks in @BENCH, panoptic segmentation and image captioning. To include token-level output like VQA or OCR, X-Decoder requires a specific head for each task. This paradigm (in Fig 3) will make the structure of the entire model very bloated, which is a major drawback for portable assistance systems. In contrast, we use task-specific prompt to build a unified input paradigm "image

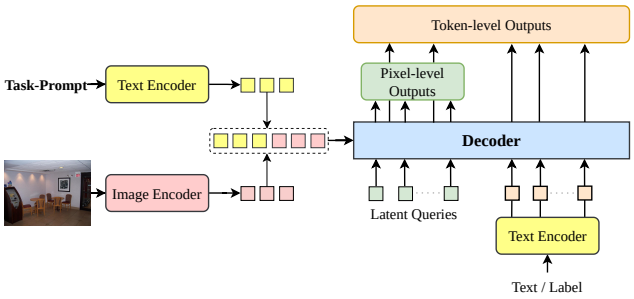

Figure 2. **Overall architecture of @MODEL.** We propose task-based prompts to unify inputs and perform different tasks all at once.

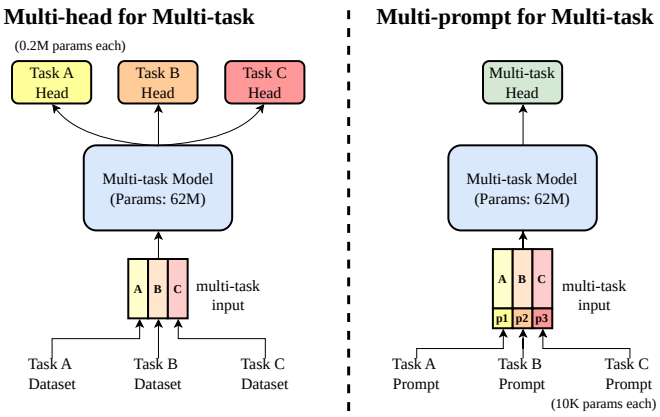

Figure 3. **Paradigms of multi-task methods.** Our @MODEL incorporates task-specific prompts that effectively unify tasks all at once and with almost no additional parameters.

+ prompt" as shown in Fig. 3. Compared with multi-head output, the benefits are three-fold: (1) Unifying input forms for different tasks. For example, it can unify the inputs of VQA (image and question text) and segmentation (image) in the manner of "image + prompt". (2) Enabling the model to distinguish different tasks, extract corresponding features in early phase. (3) Reducing the number of parameters. We design a corresponding prompt shown in Fig. 1 for each task. For VQA, we use its own questions directly as the prompt. In Sec. 5.5, we analyze the performance and efficiency advantages of such a design.

Character-based Tokenizer with Limited Vocabulary for OCR. In the text encoders of @MODEL, we use pre-trained CLIP [31] subword-based tokenizer with a vocabulary containing approximately 50,000 subwords as default. But during language-related tasks training, we need to consider a problem, there is a mismatch between dataset's vocabulary space and model's prediction vocabulary space. For example, the dataset of captioning contains a rich vocabulary that is comparable to the vocabulary that can be predicted by the model. But for OCR, we have some observations: (1) In the English OCR datasets in @BENCH, an image usually contains only one pure text. (2) The texts basically use 26 English letters and 10 numbers. (3) If each text is divided by a single character, the length is basically less than 15. When we use the default subword-based tokenizer and complete vocabulary, we found that the training effect is unsatisfactory. Subword-based tokenizer with a large vocabulary that can provide semantic information is not effective enough for OCR. The model only needs to recognize the text but not the representations of the text. A character-based tokenizer can bring a much smaller vocabulary and relieve this mismatch. More details locate in the supplementary material. The respective experiment is presented in Sec. 5.5.

Method	*PS* ADE-150 PQ	*DE* NYU-V2 RMSE ↓	*OCR* 6 Datasets avg Acc(%)	*IC* VizWiz_Cap B@1	*IC* VizWiz_Cap CIDEr	*VQA* VizWiz_VQA Acc(%)	#Params
Unified-IO (S)†	–	0.649	–	⋆	⋆	42.4	71M
Unified-IO (B)†	–	0.469	–	⋆	⋆	45.8	241M
Unified-IO (L)†	–	0.402	–	⋆	⋆	47.7	776M
X-Decoder (T)†	41.6	–	–	⋆	⋆	–	164M
GIT†	–	–	–	⋆	113.1	68.0	0.7B
PaLI†	–	–	–	⋆	117.2	67.5	3.0B
Ours	38.5	0.425	80.1	61.0	52.5	53.7	62M

Table 3. **Comparison of multi-task training @MODEL and other generalist models.** We report the multi-task training results without any pre-training and task-specific fine-tuning. Note: GIT and PaLI are LVLMs. "⋆" denotes the model has the capability for the task but does not have number reported. "–" means the model does not have the ability for the specific task. "†" means the model uses pre-trained weights for training. (B@1 = BLEU-1)

5. Experiments

This section provides experimental results and analyses to demonstrate the effectiveness of our proposed model. Implementation details are in the supplementary material.

5.1. Comparison with Existing Generalist Models

In Table 3, we compare the performances of @MODEL and some other generalist models on the tasks in @BENCH. @MODEL is the first generalist model to support assistance-related vision-language tasks and can achieve superior results. However, (1) most generalist models only include a few specific tasks and do not include all tasks in @BENCH, especially OCR task. And (2) most generalist models only report fine-tuned results and few methods report their results on assistance-related datasets. Therefore, we only compare a few related methods. Since generalist models aim to process different tasks with shared architecture and parameters, some generalist models only report results with task-specific fine-tuning, *e.g.*, X-Decoder, GIT [40] and PaLI, this fine-tuning will lose the general modeling ability, so we

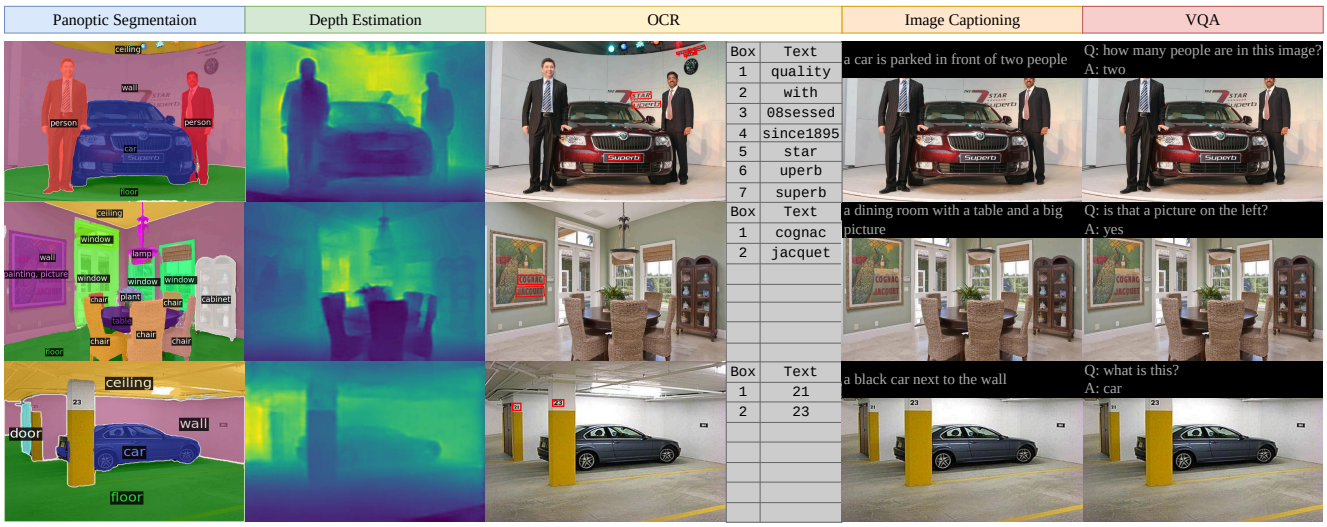

Figure 4. **Examples of multi-task training results on 5 tasks.** Given one image as input our @MODEL can output all predictions.

report the numbers without any task-specific adaptation in the manner of multi-task training. Nonetheless, with similar number of parameters, @MODEL can outperform Unified-IO (S) in depth estimation and VQA by 0.224 and 11.3%, respectively, and even better than the Unified-IO (B). On other tasks, since many methods are pre-trained, fine-tuned, and have a larger number of parameters, there is still a certain gap between @MODEL and these methods. The visualization of some multi-task training results is presented in Fig. 4.

5.2. Comparison with Specialized SoTA Models

Due to the scarcity of existing VLMs capable of encompassing all five tasks of our @BENCH, to comprehensively showcase the effectiveness of our model, we compare our multi-task model with previous single-task SoTA models in Table 4. The representative works for each task are: MaskFormer [7], Mask2Former [6] and kMaX-DeepLab [44] for panoptic segmentation; BTS [21], DPT [32] and GLP [19] for depth estimation; ASTER [35], SEED [30] and MaskOCR [25] for OCR; VizWiz_Cap, AoANet [13] for captioning, VizWiz_VQA, S-VQA [18] and CS-VQA [15] for VQA. Note that all these specialized methods are tailored for their respective specific tasks, while @MODEL is proposed to cover all these tasks. @MODEL can achieve, and in some cases, even surpass the single-task SoTA model on multiple tasks. For example, @MODEL outperforms the MaskFormer (+4.5%) and gets comparable results with pre-trained models, such as Mask2Former and kMaX-DeepLab. Even for highly competitive OCR tasks, @MODEL can outperform non-pretrained SoTA models, ASTER and SEED, by 3.3% and 1.7%, respectively. These results show that our @MODEL

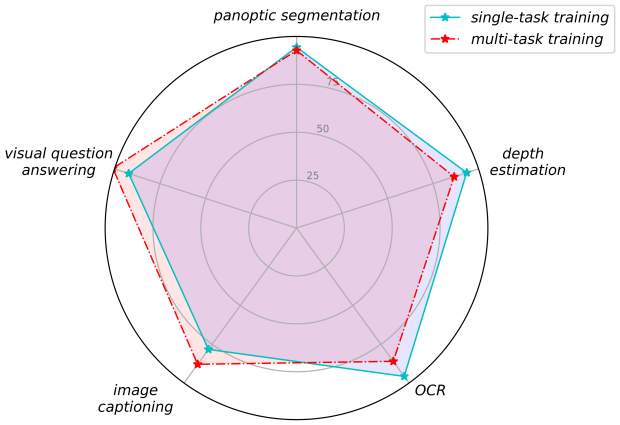

Figure 5. **Single-task and multi-task training performance** (relative) against the specialized SoTA models on different tasks

significantly bridges the performance gap between generalist models and strong baselines.

5.3. Multi-task Training *v.s.* Single-task Training

As depicted in Fig. 5, @MODEL performs less effectively in segmentation, depth estimation, and OCR under multi-task training, while demonstrating superior performance in captioning and VQA. We analyze that due to the huge OCR datasets, it is difficult to balance various tasks during multi-task training, resulting in a decline in OCR performance. For captioning and VQA, these two tasks are related to the scene understanding, they can promote each other during joint training. Furthermore, panoptic segmentation can increase the scene perception ability. See Appendix. for more analysis.

Method	*PS* ADE-150 PQ	Method	*DE* NYU-V2 RMSE↓	Method	*OCR* IC13	IC15	SVT	IIIT5K	SVTP	CUTE	avg	Method	*IC* VizWiz_Cap B@1	CIDEr	Method	*VQA* VizWiz_VQA Acc (%)
					Acc (%)											
MaskFormer† (45M)	34.7	GLP† (62M)	0.344	ASTER	91.8	76.1	89.5	93.4	78.5	79.5	86.7	VizWiz_Cap †	62.1	48.2	VizWiz_VQA	47.5
Mask2Former† (44M)	39.7	DPT*† (123M)	0.357	SEED	92.8	80.0	89.6	93.8	81.4	83.6	88.3	AoANet	65.9	59.7	S-VQA	51.6
kMaX-DeepLab† (57M)	41.5	BTS† (47M)	0.392	MaskOCR† (97M)	98.1	87.3	94.7	95.8	89.9	89.2	93.1	–	–	–	CS-VQA	53.2
Ours (62M)	39.2	Ours (62M)	0.386	Ours (62M)	97.1	84.4	92.6	90.3	88.7	93.1	90.0	Ours (62M)	60.0	45.1	Ours (62M)	49.1

Table 4. **Comparison of single-task training @MODEL and specialized SoTA models.** Note: "model (#params)" donates the number of parameters of the model. DPT* is trained with an extra dataset.

Method	Type	*PS* ADE-150 PQ	*DE* NYU-V2 RMSE ↓	*OCR* IC13	IC15	SVT	IIIT5K	SVTP	CUTE	avg	*IC* VizWiz_Cap B@1	CIDEr	*VQA* VizWiz_VQA Acc (%)	#Params
				Acc (%)										
X-Decoder (our impl.)	original	37.7	–	–	–	–	–	–	–	–	57.8	46.8	–	62M
X-Decoder (our impl.)	multi-head	38.1	0.432	89.6	68.3	80.5	84.4	73.0	77.1	79.4	59.5	50.0	–	63M
Ours	task-prompt	**38.5**	**0.425**	**90.2**	**68.7**	**81.1**	**84.9**	**73.8**	**77.8**	**80.1**	**61.0**	**52.5**	**53.7**	62M

Table 5. **Ablation study of task-specific prompt** in @MODEL. "our impl." means our implement based on the original paper, "multi-head" means we add multiple output heads (a 3-layer MLP for each task) to the original model to achieve different tasks on @BENCH.

5.4. Efficiency-Performance Trade-off

@MODEL exhibits a favorable balance between efficiency and performance. As demonstrated in Table 3, @MODEL and the small version of Unified-IO have a similar number of parameters, yet @MODEL outperforms Unified-IO (S) in all tasks they have in common, and surpass the basic version of Unified-IO. In Table 4, with almost the same number of parameters, the performance of @MODEL is better than many single-task SoTA models without pre-training. This further demonstrates that @MODEL can achieve the SoTA level on multiple tasks with fewer parameters, which is crucial for ATs with limited computational capacity.

5.5. Ablation Study

To dive deep into the model design and to investigate the effect of the proposed method, we conduct ablation studies of the task-specific prompt, the tokenizer and vocabularies.

From Task-specific Prompt to Unified Input. As shown in Table 5, the "image + prompt" training paradigm is demonstrated to be more concise and compatible with a wider range of tasks, outperforming multi-output heads training across all tasks. While there isn't a significant difference in the number of parameters between the two approaches, one potential advantage of prompt-based training is its ability to use prompts to differentiate between various tasks during training earlier. This enables the model to extract distinct features for each task, leading to varied and improved results. During multi-output head training, models extract mixed features, and only the final output head is used to decode corresponding features for different tasks. Besides, advanced output heads should be added when making the models with a multi-output head paradigm compatible with more tasks. Therefore, for generalist models, the prompt-based training paradigm is more promising.

Tokenizer sub	Tokenizer ch	Vocabulary complete	Vocabulary limited	IC13	IC15	SVT	IIIT5K	SVTP	CUTE	avg
				Acc (%)						
✓		✓		95.1	73.5	90.1	82.2	84.5	81.6	82.4
	✓	✓		95.8	80.0	90.0	87.5	85.0	90.0	86.7
	✓		✓	**97.1**	**84.4**	**92.6**	**90.3**	**88.7**	**93.1**	**90.0**

Table 6. **Ablation study of tokenizer and vocabulary** for task text recognition. "sub" and "ch" denote subword-based and character-based, respectively.

Different Tokenizers and Vocabularies for OCR. As shown in Table 6, utilizing a character-based tokenizer can lead to a performance improvement of 4.3%, and incorporating a limited vocabulary can further enhance performance by 3.3%. OCR solely recognizes image information, while captioning and VQA require processing and logical reasoning of image information to obtain reasonable answers. Therefore, when integrating tasks at different granularities, such as "relatively simple" OCR and "more complex" captioning into one model, it is suggested to carefully select different tokenizers and vocabularies to achieve the best performance on each task.

6. Discussion

Qualitative Analysis of User Study. Guided by the human-in-the-loop user study, we create the benchmark @BENCH. Score potential functions from multiple perspectives to ensure the selected functions and tasks are reasonable and important to PVIs. The five tasks in @BENCH are closely related to these functions and are one of the ways to achieve these functions. Furthermore, based on these tasks implemented by @MODEL, we are prepared to carry out further assistive function development in the future, such as indoor obstacle avoidance, text recognition, common objects detection, initial understanding of unfamiliar scenes, *etc*.

Additionally, we analyze the comments given by the blind user and list a few. About OCR, *"Usability by completely blind people would be very important to me. Existing systems of this type have difficulty selecting the right target. Text recognition on a document in front of me or on packaging in my hand is also very possible with a smartphone. However, if the function is able to read door signs, hanging posters or street signs that are not within direct reach, that would be an extremely useful function."*. According to the scores in the questionnaire and comments, PVIs attach great importance to text recognition, especially in some special scenarios. Therefore, unlike other generalist models that ignore OCR task, @BENCH contains OCR task and introduce multiple OCR datasets, covering various text scenarios. About object recognition, *"This is a function I would definitely expect!", "It would be important, on the one hand, to have a high level of reliability of recognition and, on the other hand, to be able to determine, even for completely blind people, that the correct object is being recognized."*. By introducing multi-category ADE20K and combining panoptic segmentation, @MODEL can recognize a large number of stuffs and things, compared with detection, semantic segmentation. At the same time, the more categories of objects are recognized, the accuracy of recognition will increase and the recognition will be more reliable. In sum, these positive comments and great suggestions inspire our work and provide guidance for future work of exploring VLMs for assistive technology.

Future Directions. The extensive quantitative and qualitative results have demonstrated the strong effectiveness and efficiency of our @MODEL for a variety of assistance-related tasks at different granularities. Upon the current, we see two directions worth future explorations: (1) *Pre-training*. Currently, we did not perform pre-training and @MODEL can reach a level close to the pre-trained SoTA. We believe that after pre-training, @MODEL can achieve higher performance. (2) *Functions development and deployment*. Going back to the user study, we came up with the idea for work precisely because we understood the difficulties that PVIs encounter in daily life. Existing assistive systems generally can only implement one or a few functions. A future work is to implement a multi-functional assistive system based on @MODEL. More discussions locate in the supplementary material.

7. Conclusion

In this work, we introduce @BENCH, a multi-modal, multi-task, multi-dimension benchmark for the evaluation of generalist VLMs that can empower assistive technology and help PVIs. Based on the human-in-the-loop user study with the target group, our @BENCH not only considers 5 practical tasks closely related to the daily lives of PVIs, but also takes into account the efficiency guideline for VLMs. Furthermore, we present a unified and multi-task @MODEL to address the multiple vision-language tasks. Thanks to the unified task-specific prompt design, our model can use one suit of parameters to address all 5 tasks and achieve competitive results. Extensive experiments and qualitative analysis prove the effectiveness of the proposed @BENCH and @MODEL in helping PVIs. We hope this work can provide inspiration for the design of next-generation assistive systems for helping PVIs.

Acknowledgement. This work was supported in part by the Ministry of Science, Research and the Arts of Baden-Wurttemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, in part by the Federal Ministry of Education and Research (BMBF) through a fellowship within the IFI program of the German Academic Exchange Service (DAAD), and in part by Future Mobility Grants from InnovationCampus Future Mobility (ICM). We thank HoreKA@KIT, HAICORE@KIT, and bwHPC supercomputer partitions.

References

[1] Aitor Aladren, Gonzalo López-Nicolás, Luis Puig, and Josechu J Guerrero. Navigation assistance for the visually impaired using rgb-d sensor with range expansion. *IEEE Systems Journal*, 2014. 1, 2

[2] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *CVPR*, 2023. 2

[3] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *UIST*, 2010. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 11

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6, 11

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021. 6

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 11

[9] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *CVPR*, 2019. 2

[10] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 1, 2, 4

[11] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*. Springer, 2020. 1, 2, 4

[12] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 2

[13] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 6

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 3

[15] Rachana Jayaram, Shreya Maheshwari, Hemanth C, Sathvik N Jois, and Dr. Mamatha H.R. Cross-attention with self-attention for vizwiz vqa. In *CVPR*, 2021. 6

[16] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *ICDAR*. IEEE, 2015. 4

[17] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*. IEEE, 2013. 4

[18] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 6

[19] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 6, 11

[20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 3

[21] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 6

[22] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *CVPR*, 2023. 2

[23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 3

[24] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1, 2

[25] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. Maskocr: text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022. 2, 6

[26] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*. BMVA, 2012. 4

[27] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002. 4

[29] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, 2013. 4

[30] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *CVPR*, 2020. 2, 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 2, 5

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 6

[33] Jarek Reynolds, Chandra Kanth Nagesh, and Danna Gurari. Salient object detection for images taken by people with vision impairments. *arXiv preprint arXiv:2301.05323*, 2023. 2

[34] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 2014. 4

[35] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *TPAMI*, 2018. 6

[36] Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, and Arun Sacheti. Gem: A general evaluation benchmark for multimodal tasks. *arXiv preprint arXiv:2106.09889*, 2021. 3

[37] Wu Tang, De-er Liu, Xiaoli Zhao, Zenghui Chen, and Chen Zhao. A dataset for the recognition of obstacles on blind sidewalk. *Universal Access in the Information Society*, 2023. 2

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 4

[39] Hsueh-Cheng Wang, Robert K Katzschmann, Santani Teng, Brandon Araki, Laura Giarré, and Daniela Rus. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In *ICRA*. IEEE, 2017. 2

[40] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 5

[41] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV*. IEEE, 2011. 4

[42] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*. PMLR, 2022. 1, 2

[43] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 11

[44] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *ECCV*. Springer, 2022. 6

[45] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *NeurIPS*, 2022. 2

[46] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In *ICCV*, 2021. 1, 2

[47] Xingli Zhang, Lei Liang, Shenglu Zhao, and Zhihui Wang. Grfb-unet: A new multi-scale attention network with group receptive field block for tactile paving segmentation. *Expert Systems with Applications*, 2023. 2

[48] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023. 2, 3

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 2019. 3

[50] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *arXiv preprint arXiv:2205.15237*, 2022. 3, 4

[51] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 1, 2, 4, 11

A. Model Architecture

The overall architecture of @MODEL is a generic encoder-decoder design as shown in main paper. We follow X-Decoder [51] to adapt Focal-T [43] as image encoder $\mathbf{Enc_I}$ and use a number of transformer layers as text encoder $\mathbf{Enc_T}$. Decoder is a common Transformer [51] decoder structure with self- and cross-attention layers.

A.1. Formulation

First, we use image encoder $\mathbf{Enc_I}$ to extract multi-scale features $\mathbf{Z}$ from input image $\mathbf{I}\in\mathcal{R}^{H\times W\times 3}$:

$$\mathbf{Z} = \mathbf{Enc_I}(\mathbf{I}) = \langle \mathbf{z}_l \rangle_{l=1}^{L} \tag{1}$$

where $\mathbf{z}_l \in \mathcal{R}^{H_l\times W_l\times d}$ and $\{H_l, W_l\}$ is the size of feature map at level l and d is the feature dimension. Then, we use the text encoder $\mathbf{Enc_T}$ to encode a task-specific prompt into $\mathbf{P} = \langle p_1, \cdots, p_n \rangle$ of length n. Afterwards, we use the same text encoder $\mathbf{Enc_T}$ to encode a textual label into $\mathbf{Q^t} = \langle {q_1}^t, \cdots, {q_n}^t \rangle$ and create a latent queries $\mathbf{Q^l} = \langle {q_1}^l, \cdots, {q_m}^l \rangle$ as inputs of decoder. All these features are fed into @MODEL to predict the outputs:

$$\langle \mathbf{O^p}, \mathbf{O^s} \rangle = \mathbf{@Model}(\langle \mathbf{P}, \mathbf{Z} \rangle ; \langle \mathbf{Q^l}, \mathbf{Q^t} \rangle), \tag{2}$$

where $\mathbf{O^p}$ and $\mathbf{O^s}$ are the pixel-level outputs and token-level semantic outputs, respectively.

A.2. Tasks

Based on the aforementioned designs, @MODEL can be effectively employed to integrate various vision and vision-language tasks by utilizing different input combinations.

Pixel-level Output Tasks. For these tasks, such as panoptic segmentation and depth estimation, there is no textual label as input for decoder:

$$\mathbf{O^p} = \mathbf{@Model}(\langle \mathbf{P}, \mathbf{Z} \rangle ; \mathbf{Q^l}), \tag{3}$$

where $\mathbf{O^p}$ has the same size of $\mathbf{Q^l}$.

Token-level Output Tasks. For OCR, captioning and VQA, they require both latent and text queries as inputs. Hence, Eq. (2) is adapted to:

$$\mathbf{O^s} = \mathbf{@Model}(\langle \mathbf{P}, \mathbf{Z} \rangle ; \langle \mathbf{Q^l}, \mathbf{Q^t} \rangle), \tag{4}$$

where $\mathbf{O^s}$ correspondingly has equal size of $\mathbf{Q^t}$, and no pixel-level output are predicted. All predictions follow an auto-regressive strategy.

B. Loss Functions

B.1. Pixel-level Output Loss

Segmentation Loss. There are two losses on the segmentation corresponding to two tasks. For mask classification, we use text encoder $\mathbf{Enc_T}$ to encode all N class names including "background" into N text embeddings $\mathbf{E}_{cls}\in\mathcal{R}^{N\times C}$ and take it to represent the concept. Afterward, we take the first $(m-1)$ latent queries and compute the dot-product between these outputs and concept embeddings to obtain an affinity matrix $\mathbf{S}_{cls}\in\mathcal{R}^{(m-1)\times N}$ and compute $\mathcal{L}_{cls}$=$\mathbf{CE}(\mathbf{S}_{cls}, \mathbf{y}_{cls})$, with the ground-truth class $\mathbf{y}_{cls}$. For mask prediction, we use Hungarian matching [4, 6] to find the matched entries of first $(m-1)$ outputs to ground-truth annotations. Afterward, we to use binary cross-entropy loss $\mathcal{L}_{bce}$ and dice loss $\mathcal{L}_{dice}$ to compute the loss. Thus, the overall training loss function of panoptic segmentation is:

$$\mathcal{L}_{ps} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{5}$$

where λ_{cls}, λ_{bce} and λ_{dice} are coefficient weights to control different losses

Depth Estimation Loss. Given the prediction $\mathbf{O^p}$ derived from m latent queries, we use the last (m-th) latent querie to make depth prediction. In order to calculate the distance between predicted output $\widehat{\mathbf{Y}}_{de}$ and ground truth $\mathbf{Y}_{de}$, we use scale-invariant log scale loss [8, 19]. The equation of training loss is as follows:

$$\mathcal{L}_{de} = \frac{1}{n}\sum_i {d_i}^2 - \frac{1}{2}(\frac{1}{n}\sum_i d_i)^2, \tag{6}$$

where $d_i = log(y_i) - log(\widehat{y}_i)$, y_i and $\widehat{y}_i$ are ith pixel-value of $\mathbf{Y}_{de}$ and $\widehat{\mathbf{Y}}_{de}$, respectively.

B.2. Token-level Output Loss

For token-level tasks, we begin by extracting embeddings for all tokens in the vocabulary, which has a size of V, from the text encoder. Using the last n semantic token-level outputs from @MODEL, we calculate the dot product with all token embeddings to generate an affinity matrix $\mathbf{S}_{token}\in\mathcal{R}^{n\times V}$. Subsequently, we compute the cross-entropy loss $\mathcal{L}_{token}$=$\mathbf{CE}(\mathbf{S}_{token}, \mathbf{y}_{token})$, where $\mathbf{y}_{token}$ represents the ground-truth next-token id.

B.3. Multi-task Training Loss

During multi-task training, we calculate losses on the top decoder layers for each task to guide the model to converge faster in the early training stage and accelerate the overall training process. The overall training loss function is:

$$\sum_{task\in\{ps,de,ocr,ic,vqa\}} \sum_{i=1}^{nl_{task}} \lambda_{task}\mathcal{L}_{task}, \tag{7}$$

where nl_{task} represents the number of decoder layers that need to calculate the loss for different task, λ_{task} and $\mathcal{L}_{task}$ are loss weights and losses for different task, respectively.

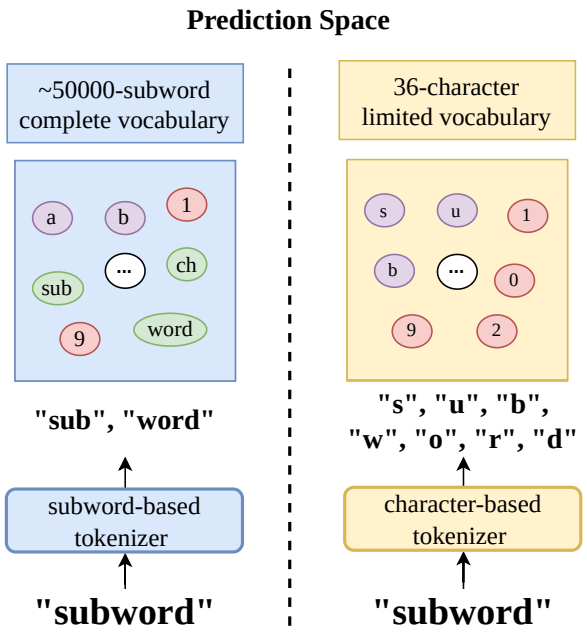

Figure 6. Comparison between subword-based tokenizer and character-based word tokenizer in our proposed @MODEL

C. Implementation Details

C.1. Multi-task Training

Training Setting. Since the number of images in OCR training dataset is much larger than datasets for the other tasks, we define OCR training dataset as the major dataset for multi-task training. It means that the total number of iterations is calculated based on the number of images in the OCR dataset. The batch sizes for panoptic segmentation, depth estimation, OCR, captioning, and VQA are 4, 4, 768, 8, and 4, respectively, to accommodate datasets of different sizes for various tasks. The model is trained with 15 epochs based on OCR datasets on 4 A100 (40G). The AdamW optimizer is used with the initial learning rate 1^e-5. A step-wise scheduler is used to decay the learning rate by 0.1 on the fraction [0.6, 0.8] of training steps.

Hyperparameter Choice. In @MODEL, the decoder has 7 decoder layers. Due to segmentation and OCR are the top two scored tasks (from user study), and the OCR datasets are very large, the amount of data for each task is unbalanced, we set nl_{task} as 6, 3, 6, 3 and 3 for panoptic segmentation, depth estimation, OCR, captioning and VQA, respectively, to allow the model to focus more on segmentation and OCR during training. We set loss weights λ_{task} as 1, 10, 10, 2 and 2, respectively. Because the loss values of OCR and depth estimation in the later stage of training are very small, in order to minimize significant differences in the loss magnitude for each task as much as possible, we have made such a setting. And in Eq. (5), we set λ_{cls} = 2, λ_{bce} = 5 and λ_{dice} = 5 as default, following the settings of X-Decoder.

C.2. Single-task Training

All tasks are trained with AdamW as the optimizer on 4 1080Ti (11G), except OCR. The initial learning rate is 1^e-5 and reduced by 10 times after 60% and 80%.

Panoptic Segmentation. We train the model for 50 epochs. We set the image resolution to 640×640 and the batch size to 4.

Depth Estimation. We train the model for 50 epochs. We set the image resolution to 480×640 and the batch size to 16. Note that this task is very unstable and requires careful hyperparameter tuning. If you encounter training errors, you can increase the batch size, reduce the learning rate or training with single precision (FP32).

Optical Character Recognition. We train the model for 10 epochs on 4 A100 (40G). We set the image resolution to 64×200 and the batch size to 1024.

Image Captioning. We train the model for 50 epochs. We set the image resolution to 480×640 and the batch size to 16. We use all captions for training and do not use beam search and CIDEr optimization.

Visual Question Answering. We train the model for 30 epochs. We set the image resolution to 480×640 and the batch size to 16.

C.3. Character-based Tokenizer with Limited Vocabulary for OCR

In our main paper, we observed that subword-based tokenizer with complete vocabulary hurts the performance of OCR task. In Fig. 6, we show how to use character-based tokenizer and much smaller limited vocabulary to perform OCR task. Using a character-based word tokenizer to divide the text that needs to be recognized into characters one by one, model only needs to predict token from the limited vocabulary space, and do not need to select candidate subword from the complete vocabulary. This reduces the prediction space and improves the accuracy of prediction.

D. User Study

D.1. Comments on Generalist Assistance Systems

By conducting the questionnaire survey, we communicated with visually impaired individuals to comprehend the functionalities they expect a generalist assistive system should possess. We got some thoughts like: *"It should find the door, look for stairs in an open area, read the*

house/room number, read signs/plates, describe the environment, warn me of obstacles, and can navigate the corridors with a floor plan.". Some participants also described specific usage scenarios, *"I would use navigation and obstacle detection systems outside. It should warn me of obstacles or describe something I'm about to encounter. For example, if I'm navigating outside and there's a road ahead, then it should say if it has a roundabout or an intersection. Or, if there is a railroad crossing, announce something similar. It would be cool if there was an all around view. The system says, the front of you is street and the back is a building, left is bike racks, etc. If there is a name of the store, read it out. The most important thing is to have a general navigation ability based on this all around view. If I then say navigate to the store (name of the store) recognized by this system, then it should navigate me there."*. Based on these thoughts and comments, essential functions identified by People with Visual Impairments (PVIs) that a generalist assistive system should include are:

(1) **Navigation and Obstacle Avoidance.** A critical component is a navigation system integrated with obstacle detection capabilities. PVIs desire a system that allow for interactive navigation, where users can request directions to specific locations identified by the system.
(2) **Text Recognition and Environmental Description.** The ability to recognize and verbally relay textual information is also important. This includes identifying and reading door labels, room numbers, and signs. Furthermore, recognizing the names of stores, significant landmarks or other text contributes to better environmental understanding and orientation.
(3) **Comprehensive Scene Interpretation.** PVIs expressed a desire for a system that provides a holistic view of their surroundings. This "all-around vision" function should describe streets, buildings, and other elements in the vicinity.
(4) **Integration of Text-to-Speech Technology.** Incorporating text-to-speech technology for dynamic interaction is also valuable.

D.2. More Comments

Navigation. The majority of participants (5 P: 5 participants) prioritize outdoor navigation, noting its greater complexity and risk. They highlight that outdoor environments pose larger obstacles, longer and more complex routes, and a higher likelihood of getting lost compared to indoor scenarios. One participant emphasized, *"Outdoor navigation is much more important. Indoors, the reach of a cane is much more likely to adequately capture the surroundings. The distances are shorter and the density of people is higher."*. Another added, *"Definitely outdoors. If I have to go into a building I don't know, it will probably only be for once. It's not worth learning a way to do that."*. The unpredictability of outdoor spaces, such as traffic, was also mentioned as a significant factor. Conversely, a minority (2 P) believes that indoor navigation is more important. They mention the challenges of navigating within large unfamiliar buildings, locating specific rooms, stairs, elevators, or exits, walking across a large open-area, and walking in rooms with highly differentiated structures, such as restrooms. Importantly, they spend most of their time indoors.

Text Recognition. Today, PVIs mainly use screen readers to recognize digital texts and usually use smartphones, or smartphones Apps to read non-digital texts. However, they find non-digital text reading is difficult and cumbersome, like *"Everything I receive on paper in the post annoys me. I use apps like Seeing AI and Be My Eyes or the iPhone's magnifying glass to read non-digital texts, but using a smart glass to read these text directly would be better."*. They also pointed out that it is also important for them to read signs to find the right floor or hallway and read door numbers to enter the right room.

Other Functions. About depth estimation, *"This function helps one develop a mental map of an environment. You get the proportions well."*. About object location, *"In my personal environment I am always very sure where all the things I am looking for are. However, locating a true one in a larger shelf section of 3-4 meters would be very useful. A function that detects objects that don't belong in that space would also be very helpful to check a room for overlooked clutter. The glasses could use a reference photo of the tidy room and then report any anomalies, such as dirty dishes on the table or socks on the floor."* and *"If I only need it if I can't find something in my apartment, it could make the search easier, but I would need it pretty rarely."*. About surroundings understanding, *"It would be important to me that the description be highly efficient. The short form is always first"* and *"Most of the time we are not interested in the scene because it is too much information for us. But descriptions of photos, environments, etc. are very exciting. ChatGPT is really great."*. About scene recognition, *"Perhaps interesting for recognize different scenarios, but a correspondingly efficient description of the image would serve the same purpose. I can't imagine a situation where I would need room detection. I usually know which room I'm going to or being led into."*. About visual Q&A, *"This function would make it possible to expand a short initial description of an image dynamically and according to your own needs. That would improve the overall function enormously."*.

Interaction. If there were such a general system, PVIs prefer interacting with system through discrete button presses or subtle gestures (6 P), rather than voice commands (1 P) for privacy reasons, when inputting instructions. For receiving system feedback, they show a preference for auditory feedback (for general purpose) and vibrations (for special

Task			ADE-150	VizWiz_Cap		VizWiz_VQA				
PS	IC	VQA	PQ	B@1	CIDEr	Other	Unans	Yes/No	Number	Acc(%)
✓			39.2	–	–	–	–	–	–	–
	✓		–	60.0	45.1	–	–	–	–	–
		✓	–	–	–	30.5	92.1	70.1	13.7	49.1
✓	✓		37.7	57.8	46.8	–	–	–	–	–
	✓	✓	–	59.8	46.3	32.2	86.5	73.4	16.4	48.8
✓	✓	✓	38.5	61.0	52.5	39.4	88.2	70.1	10.8	53.7

Table 7. **Comparison of results of mixed training for different tasks.** Note:"Other", "Unanswerable", "Yes/No", "Number" are 4 different answer types for VQA. (PS = panoptic segmentation, IC = image captioning).

purpose such as obstacle avoidance).

Based on these comments and ideas, it becomes evident that for PVIs, navigation and quick, direct recognition of non-digital text are the two most critical functionalities. Meanwhile, the multifaceted nature of navigation encompasses functions like environmental comprehension, obstacle avoidance, path planning, voice guidance and *etc.* These insights serve as valuable guidance for our work. Furthermore, the analysis of participants' relevant feedback has provided us with an initial understanding of creating a universal assistive system.

E. More Experiments

E.1. Complementariness in Multitasking

As shown in the experiments section, our @MODEL exhibits a strong performance in captioning and VQA under multi-task training. Here, we further study the role of segmentation objectives in vision-language (VL) understanding, as well as the role of different vision-language understanding tasks on each other. To investigate, we mix different tasks for training. In Table 7, for captioning, when jointly trained with VQA or PS, or all tasks, CIDEr improved by 1.2, 1.7 and 7.4 respectively. For VQA, we report 5 numbers for better analysis, namely the accuracy for 4 Q&A types: *other*, *unanswerable*, *yes/no*, *number*, and the overall accuracy. From the comparison of these numbers, when training VQA alone, the model tends to predict "unanswerable" to improve the accuracy. Because in the dataset, the *unanswerable* type of Q&A is the most common. For other types of Q&A, the accuracy is relatively lower because a deeper or more granular understanding of the semantic information of image is required to predict the correct answers. After joint training with captioning, the accuracy of *unanswerable* type Q&A decreased, and the accuracy of other types increased. The model does not just return "unanswerable" blindly but understands more semantic information of the image and then make predictions. When all tasks are trained together, the accuracy of *other* type Q&A is greatly improved (+8.9%). We analyze that it is because the question of this type of Q&A is usually "*what is this?*", and the segmentation task naturally has a very good assisting effect in answering this question. Segmentation data can help models to learn more fine-grained visual understanding and consequently benefit vision-language tasks. We also give some examples to show these improvements in Fig. 7. Along with our findings in the main paper, we conclude that segmentation has clear benefits to VL learning and different VL tasks are complementary to each other.

F. More Visualization

F.1. Visualization on Test Datasets

We present a comprehensive visualization of our model's performance on the test datasets in Fig. 8. For segmentation, we show some results in outdoor scene, indoor scene, multi-person scene, especially the open-area mentioned by the PVIs. For OCR, various types of text recognition results can show the robustness and generalization of @MODEL. For other task, @MODEL can also perform well.

F.2. Zero Shot

Finally, we apply the 5 tasks in a zero-shot manner to show the generalization ability of @MODEL. @MODEL performs well on three tasks: segmentation, depth estimation, and OCR, as shown in Fig. 9. However, for open-ended tasks, captioning and VQA , the performance on out-of-dataset data can sometimes be less satisfactory (Fig. 9 (B)). Therefore, it may be necessary to perform large-scale pre-training to enhance the model's capability for handling these tasks well in zero-shot.

G. More Discussion

This section discusses the limitations and future work of this work for more insights on the research in this track.

Pre-training. In the main paper, we did not perform pre-training. This has a certain impact on the capability of zero shot, especially for open-ended tasks. In the future, we plan to conduct pre-training on large-scale corpora to enhance the model's zero-shot capability. Additionally, we use a unified language encoder to encode text in @MODEL. Pre-training can enrich the vocabulary size, thereby improving the model's ability to open-vocabulary segmentation. The importance of this open-vocabulary capability for practical applications is self-evident, especially for blind users. As mentioned by blind users in user study, they require systems with high object recognition accuracy. When the model has seen a greater variety of objects and can distinguish between them, the recognition accuracy also increases. Additionally, this open-vocabulary capability allows the model to handle previously unseen objects. In sum, after pre-training, the model can better handle the diversity, complexity and unpredictability of usage scenarios.

Multi-task Training. As shown in the main submission, @MODEL performs well on the OCR task during single-

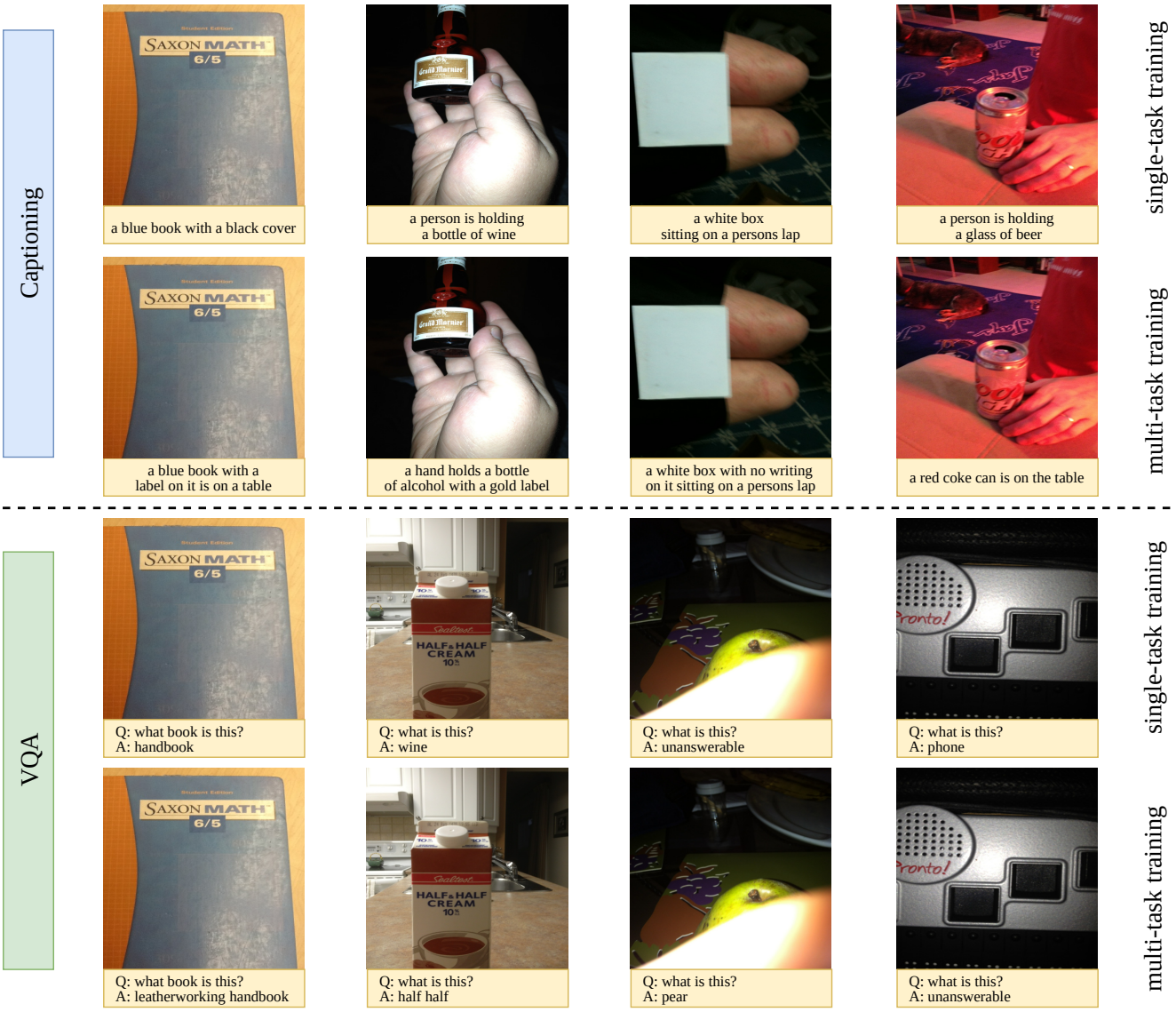

Figure 7. Examples to show the promotion of vision for vision-language and complementariness between different vision-language tasks.

task training, but there is a certain gap in performance during multi-task training. Our analysis suggests that the OCR dataset is too large, and the model does not balance multiple tasks during training. When dealing with multi-task training with extremely imbalanced dataset sizes, it is not enough to merely adjust loss weights differently. In the future, we may try more optimization methods for multi-task learning to ensure performance without greatly increasing the training time.

Functions Development and Model Deployment. In our user study, we have identified several potential and crucial functions that received unanimous agreement from participants. Furthermore, it's important to note that @MODEL is not limited to these five tasks alone; it can be extended to more uni-modal or multi-modal tasks to provide more functionalities. Our future research direction will focus on building a PVIs-Centred generalist assistive system, leveraging @BENCH and @MODEL as cornerstones, to develop a wide range of practical functions and services. As for model deployment, although @MODEL achieves high performance on multiple datasets, since the model is based on Transformer, its costs are larger than the non-Transformer models. Additionally, though @MODEL only has 62M parameters, it is still difficult to deploy such a model in the portable device used by PVIs. Therefore, in our future work we will discover how to extract or compress @MODEL into an efficient light-weight model.

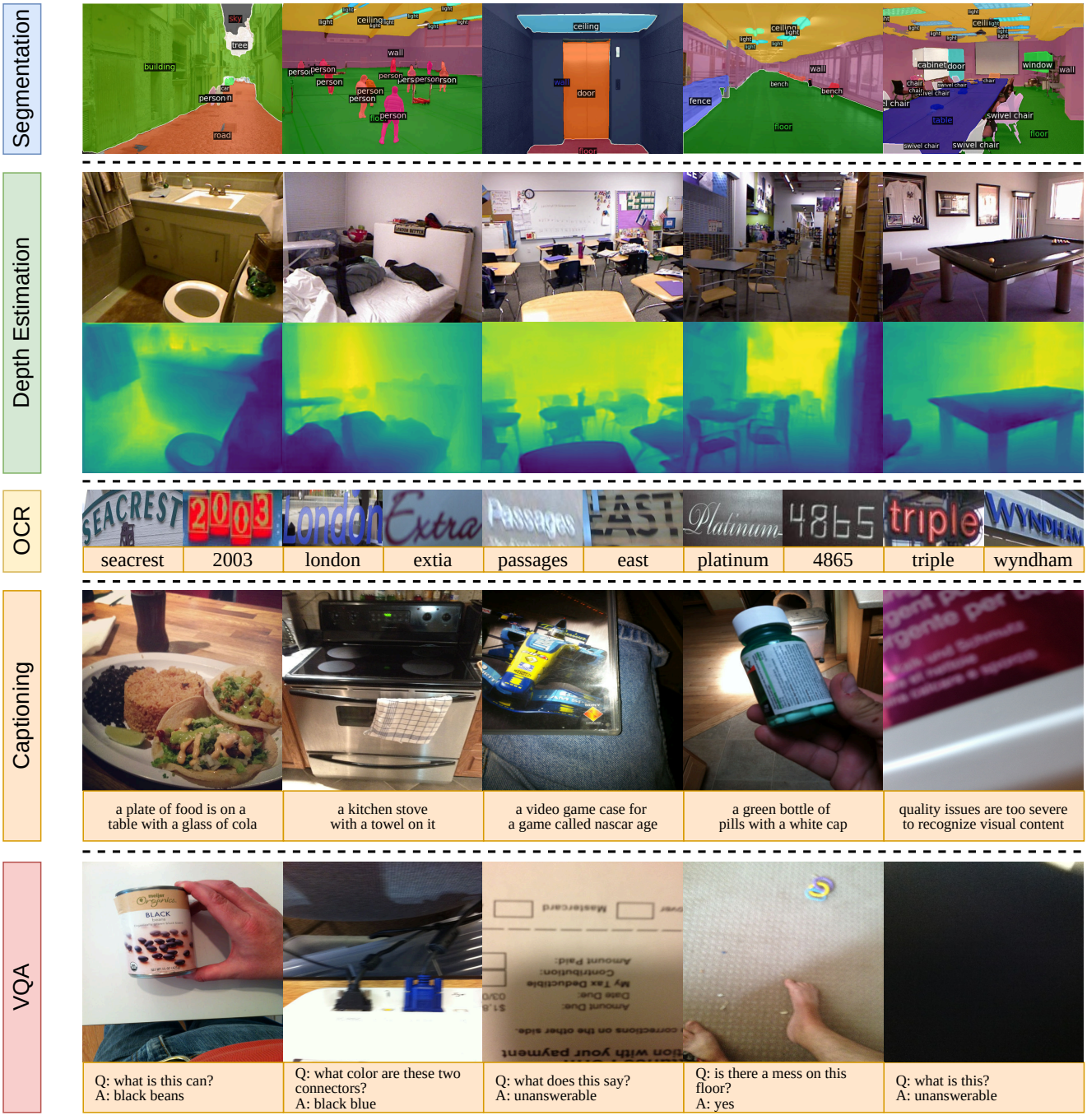

Figure 8. Examples on different test datasets. These images cover a diversity of visual domains and concepts in the daily life of PVIs.

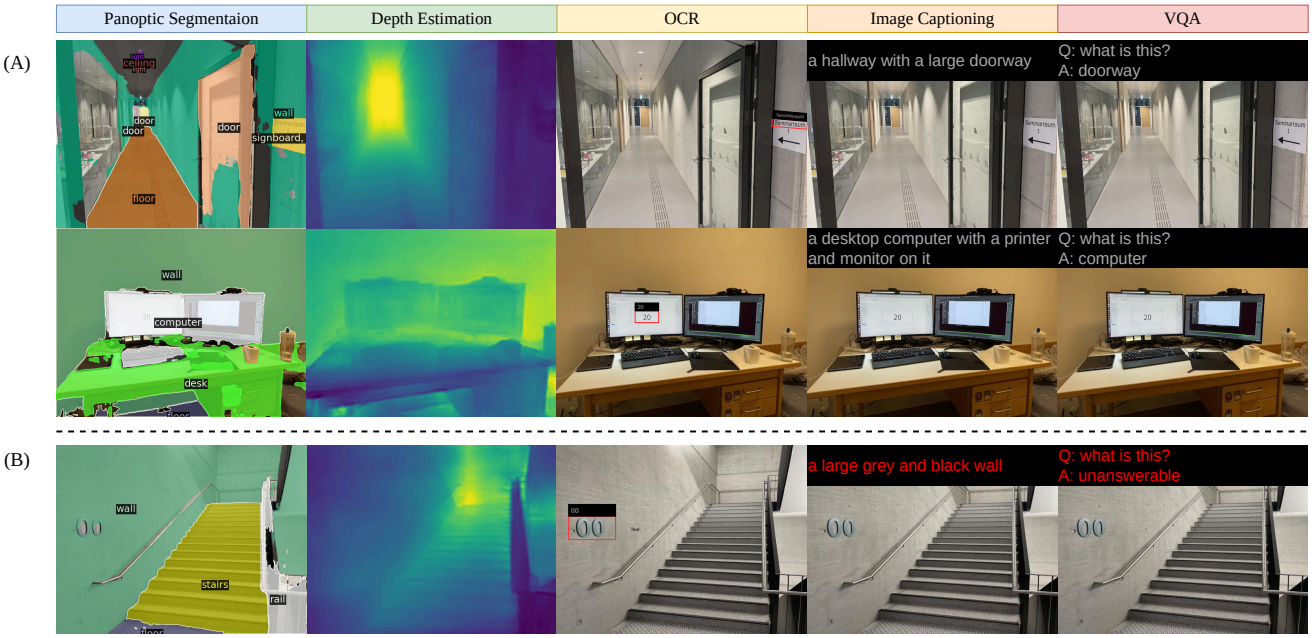

Figure 9. Examples on real-world scenes. These images were randomly collected by using mobile phone.