We appreciate the reviewers' recognition and thoughtful feedback. Below, we address the concerns raised. Dynamic query and efficiency are vital directions for future research.

Q1: Innovations and Contributions [hCDc, NYvM]. We seek to bridge the gap between word-based and sentence-based perception within the MLLM framework. Our primary contributions lie in **unleashing the generalizability of MLLMs for perception tasks through carefully designed training strategies**. Specifically: (1) We introduce a multi-granularity decoder to the MLLM framework, enabling it to predict both bounding boxes and masks, thereby equipping MLLMs with perception capabilities. (2) We maximize the use of available annotated data from multiple tasks, enriching the vocabulary and semantic diversity of the training set to better support MLLM training. (3) We propose a unified prompt template that is applicable to various tasks, facilitating joint training within a single framework. The CoT-inspired data construction is designed to further stimulate the reasoning abilities of LLMs.

Q2: Source of Improvements and Generalizability [hcDc, NYvM, a4tA]. We attribute the enhancements primarily to **joint training with dedicated prompts**, for the following reasons: (1) The complementarity between different tasks enhances the model's perceptual capabilities—detection enriches vocabulary, while segmentation provides fine-grained localization cues. (2) Our response structure, "describe-then-segment", emulates human cognition, encouraging the model to understand the image before performing segmentation. (3) Employing diverse prompts for the same training instance promotes stronger generalization and robustness. **MVP-LM achieves improvements across in-domain and out-of-domain segmentation benchmarks, as shown in Tab. 3 in our manuscript**, highlighting the strong performance and generalizability.

Q3: Decoder Details [hCDc, NYvM, a4tA]. Our decoder employs a transformer-based architecture. Initially, a set of content queries and their corresponding reference points are generated through a query selection mechanism (as seen in Fig.2 in our manuscript). These queries are then iteratively cross-attended with multi-scale visual features via deformable attention layers. The outputs from each MS-Deform layer are subsequently processed by three shared heads—a cross-modal similarity computation head, a box head, and a mask head—to produce predictions. During training, predictions from each layer are assigned to their corresponding annotations to calculate losses through Hungarian matching. However, during inference, only the final layer's output is utilized. Moreover, a denoising strategy is incorporated, similar to that used in MaskDINO, to stabilize optimization and accelerate convergence.

Q4: Training Recipe Details [hcDc, NYvM, a4tA]. Our two-stage training is based on a COCO-pretrained perception model and open-source LLM weights. Training recipes are listed in Tab. R1. Specifically, the number of steps for the ablation study is reduced to 9k. **Query numbers and Loss weights follow prior works without tuning**. Besides, the ablation study about query number can be seen in Tab. 7 in the manuscript.

Table R1. Hyperparameters for both training stages. "CP", "RC", "O", and "GG" denote COCO-Panoptic, RefCOCO, Objects365, and GoldG, respectively.

Parameters	Stage1	Stage2
Training Components	Connector	Connector + LLM + Multi-granularity Decoder
Optimizer	AdamW	
Training Rate	2×10^{-3}	4×10^{-5}
Batch Size	128	64
Number of Steps	4650	80000
Learning Rate Schedule	Cosine Decay	
Weight Decay	0.0	0.05
Warmup Ratio	0.03	
Training Data	CC3M	CP(33.3%) RC(33.3%) O(16.7%) GG(16.7%)
Loss	L_{LLM}	$L_{LLM} + 2 \cdot L_{word/sent} + 5 \cdot L_{L1} + 2 \cdot L_{GIoU} + 5 \cdot L_{BCE} + 5 \cdot L_{DICE}$
Image Size	1024×1024	
Image Processing	Resize longer to 1024 and pad shorter to 1024	

Q5: REC Metrics [hcDc]. In contrast to the enhanced multi-modality chat models like MiniGPT-v2 and Octopus, MVP-LLM is a unified framework for various perception tasks containing LLM. As shown in Tab. R2, MVP-LM outperforms the listed 7B models. REC metrics will be included in the final version.

Table R2. Comparison of REC metrics (Acc@0.5) on RefCOCO series. "Params", "Res", "R", "Rp", "Rg", "v", "tA", "tB", and "t" denote parameter size, resolution, RefCOCO, RefCOCOplus, RefCOCOg, val, testA, testB, and test, respectively.

Method	Params / Res	R(v)	R(tA)	R(tB)	Rp(v)	Rp(tB)	Rg(v)	Rg(t)
MiniGPTv2	7B / 448	88.7	91.7	85.3	79.9	74.5	84.4	84.7
Octopus	7B / 336	89.0	92.6	83.4	83.5	76.0	84.4	86.2
MVP-LM	**1.3B / 384**	**93.5**	**94.5**	**91.6**	**84.9**	**79.3**	**86.7**	**87.4**

Q6: Multi-scale Feature [NYvM]. A clear performance decline is witnessed on PQ for COCO-Panoptic (from 55.6 to 51.8) and cIoU for RefCOCO val (from 75.7 to 70.9) without multiscale feature. Averaging instead of concatenating features increases cIoU (by 0.35) but decreases PQ (by 2.41). The reason why cIoU stays stable but PQ declines sharply for different fusion operations is that (1) averaging lets small object features be overshadowed by high-level semantics, and that (2) both methods improve resolution.

Q7: Query Selection [NYvM]. Using separate learnable query embeddings, rather than applying query selection, improves RefCOCO val performance (from 75.7 to 78.3) but lowers COCO-Panoptic results (from 55.6 to 54.4). It reflects that RefCOCO's simpler targets suit fixed queries, and that COCO's complexity requires more flexible query selection.

Q8: Computational Efficiency [a4tA]. It consumes most inference time to auto-regressively generate image captions before segmentation token, but it predicts boxes and masks in a single forward pass as fast as common perception models. MVP-LM gets 0.3 QPS on RefCOCO val.