PERCEPTIONDLM: PARALLEL REGION PERCEPTION WITH MULTIMODAL DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have achieved remarkable progress in visual understanding tasks. However, most existing MLLMs rely on autoregressive generation, which limits their efficiency for perception tasks that require captioning multiple regions. In this work, we propose **PerceptionDLM**, a multimodal diffusion language model optimized for efficient parallel region perception. Built upon **PerceptionDLM-Base**, a strong foundational baseline that achieves state-of-the-art performance among open-source diffusion MLLMs, our architecture fully leverages the parallel decoding nature of diffusion language models (DLMs). Specifically, we introduce efficient prompting and structured attention masking to enable simultaneous perception of multiple masked regions, allowing the model to generate region descriptions in parallel at both the sequence and token levels. This design significantly improves inference efficiency compared with existing approaches that process regions sequentially. To systematically evaluate the parallelism property of visual perception capability for DLMs, we construct a new **Para**llel **D**etailed **L**ocalized **C**aptioning Benchmark (**ParaDLC-Bench**) by scaling DLC-Bench to include multiple region masks per image, enabling joint evaluation of both caption quality and inference efficiency. Experiments demonstrate that PerceptionDLM maintains competitive performance in region captioning while achieving substantial speed improvements for multi-region perception tasks. Our results highlight the potential of multimodal diffusion language models for efficient, parallel visual perception. To the best of our knowledge, we are the first to achieve parallel region captioning and perception by leveraging the advantages of diffusion language models. Code, models, and datasets will be released.

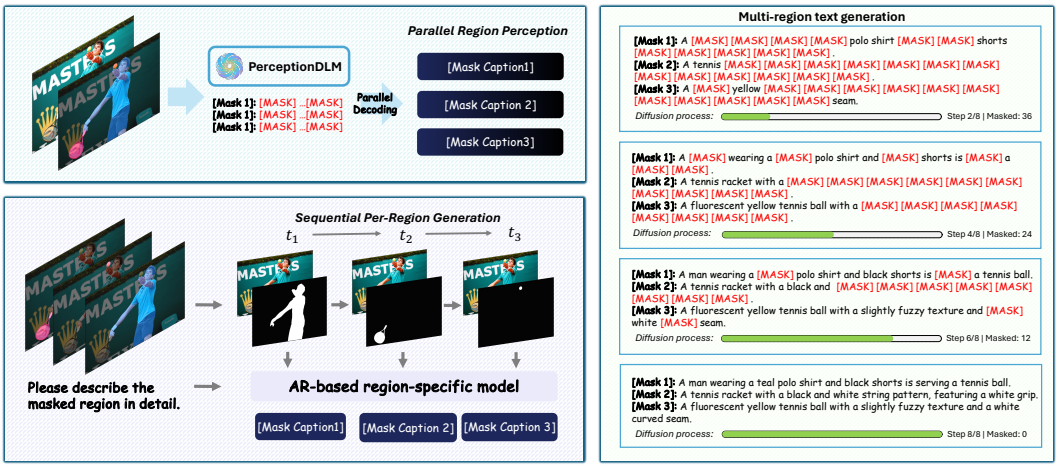
1 INTRODUCTION

Visual perception (Lin et al., 2014) is increasingly central to real-world multimodal intelligence. As applications demand more sophisticated interactions with the visual world, MLLMs have recently achieved remarkable progress across a wide range of visual understanding tasks via visual instruction-tuning and reinforcement learning pipelines (Liu et al., 2023; Li et al., 2025a; An et al., 2025; Bai et al., 2025a; Wang et al., 2025b; Zhu et al., 2025; Lei et al., 2025; Meng et al., 2025). However, while these models excel at image-level comprehension, many perception-heavy applications demand fine-grained, localized understanding. In these scenarios, a model must accurately describe multiple specific regions within a single image with fine-grained attributes and minimal confusion across targets (Rasheed et al., 2024; Lian et al., 2025; Wang et al., 2025a). Consequently, fine-grained region-level captioning has emerged as a core capability for next-generation multimodal systems.

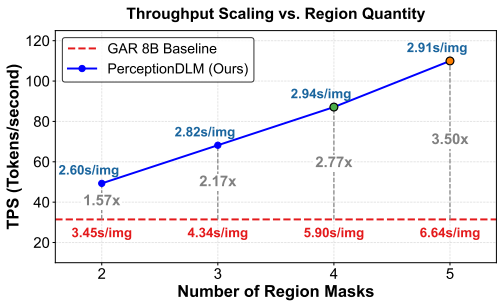
Despite high accuracy, existing MLLMs are still dominated by autoregressive (AR) decoding, which introduces an efficiency bottleneck for multi-region perception. Under the AR paradigm, descriptions are typically generated sequentially for each individual region, and each description is token-by-token (Lian et al., 2025; Wang et al., 2025a). As the number of queried regions increases, inference cost and latency grow rapidly, making the dense-region perception difficult to scale.

Recently, large diffusion language models (DLMs) have emerged as a competitive alternative to AR language modeling (Nie et al., 2025; You et al., 2025; Ye et al., 2025; Yang et al., 2025b). Their masked denoising generation paradigm enables non-autoregressive generation and exposes intrinsic token-level parallelism. Yet directly extending diffusion-based VLMs to fine-grained

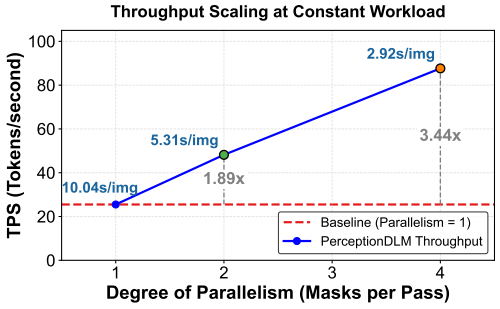
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Parallel Region Captioning with PerceptionDLM



(b) Throughput vs. Region Quantity



(c) Throughput Scaling at Constant Workload

Figure 1: **Overview and efficiency analysis of PerceptionDLM. (a)** Given an image and multiple region masks, PerceptionDLM generates captions for all regions in parallel through a denoising diffusion process. Unlike conventional AR-based models like DAM (Lian et al., 2025) and GAR (Wang et al., 2025a), which process regions sequentially, our approach avoids linear growth in inference cost. **(b)** Throughput (TPS) scaling comparison. PerceptionDLM achieves near-linear TPS growth with stable per-image latency as the number of region masks increases. **(c)** Parallelism scaling under a constant workload (4 masks per image), demonstrating up to a $3.44\times$ throughput speedup when fully parallelized.

localized perception remains non-trivial: existing multimodal diffusion language models often lack strong perceptual capabilities, and their potential for concurrent multi-region perception remains underexplored (You et al., 2025; Cheng et al., 2025; Ye et al., 2025; Yang et al., 2025b). This motivates the question of this work: *Can we design a multimodal diffusion language model that preserves strong perception quality while unlocking practical parallelism for region-conditioned captioning?*

To answer this question, we propose **PerceptionDLM**, a diffusion-based framework for parallel region perception. We first build **PerceptionDLM-Base**, a strong discrete diffusion multimodal baseline that integrates a pretrained vision encoder with a diffusion language backbone via visual instruction tuning. Building upon this foundation, we introduce region-aware structural design. As illustrated in Figure 1, this enables PerceptionDLM to generate multiple region descriptions jointly within a single denoising process rather than through independent per-region decoding passes.

To evaluate both quality and efficiency under this setting, we introduce **ParaDLC-Bench**, which extends DLC-Bench (Lian et al., 2025) from single-region evaluation to concurrent multi-region evaluation. ParaDLC-Bench explicitly measures caption quality and inference efficiency, enabling direct comparison between sequential and parallel decoding paradigms.

Extensive experiments show that PerceptionDLM provides a strong baseline among open discrete diffusion VLMs. On our 16-benchmark multimodal evaluation, PerceptionDLM-Base outperforms LLaDA-V (You et al., 2025) on 15 of 16 benchmarks. On ParaDLC-Bench, PerceptionDLM achieves an average accuracy of 62.4%, nearly doubling the performance of baselines like LLaDA-V (35.2%). Furthermore, it achieves competitive accuracy against strong AR region-specific models while substantially improving inference efficiency. As the number of queried regions increases, PerceptionDLM avoids the linear latency growth of AR models, yielding up to a $3.5\times$ throughput speedup in dense perception scenarios (e.g., 5 masks per image) at comparable model scales. Additional experiments show that this parallelism is orthogonal to step-level acceleration, which compounds with it for a further speedup. Extensive experiments show that PerceptionDLM establishes a strong state-of-the-art baseline among open discrete diffusion-based MLLMs and achieves competitive accuracy against strong AR region-specific models, while substantially improving inference efficiency under comparable model scales. Our main contributions are summarized as follows:

- We present **PerceptionDLM-Base**, a high-performance, open discrete diffusion multimodal baseline that enhances perception capabilities for diffusion VLMs.
- We propose a **parallel region perception** mechanism for DLMs that combines region-aware mask embeddings and structured attention masking to enable simultaneous multi-region caption generation.
- We build **ParaDLC-Bench**, a benchmark tailored for multi-region localized captioning that evaluates both quality and efficiency.
- We demonstrate that **PerceptionDLM** achieves a strong trade-off between accuracy and efficiency, providing competitive region-caption quality with effective inference-speed gains over sequential region-captioning pipelines, and show that these gains further compound with step-level distillation.

2 PerceptionDLM-Base: Stronger DVLM Baseline

To establish a strong and scalable baseline for diffusion-based multimodal perception, we first introduce PerceptionDLM-Base, a multimodal diffusion language model that extends large-language diffusion models to visual instruction tuning.

2.1 Diffusion Language Modeling and Multimodal Architecture

Diffusion Language Modeling Preliminaries Let $x_0 = (x_1, \ldots, x_N)$ be a sequence of discrete text tokens from a vocabulary $\mathcal{V}$. Discrete Diffusion Language Models (DLMs), such as LLaDA (Nie et al., 2025), formulate text generation as a generative Markov process. During the forward process, clean tokens x_0 are progressively corrupted into a sequence x_t at timestep $t \in (0, 1]$, typically by replacing tokens with a special absorbing state (e.g., a `[MASK]` token). The reverse process learns a neural network $p_\theta(x_0|x_t)$ to denoise x_t back to x_0. The model is optimized using a reweighted variational lower bound, which simplifies to predicting the masked tokens:

$$\mathcal{L}_{\text{DLM}} = -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t}\sum_{i=1}^{N}\mathbf{1}[x_t^i = \texttt{[MASK]}]\log p_\theta(x_0^i|x_t)\right], \tag{1}$$

where N is the sequence length and the indicator function $\mathbf{1}[\cdot]$ ensures that the loss is computed only for masked tokens.

Multimodal Architecture and Visual Instruction Tuning. To extend this diffusion language modeling paradigm to multimodal perception via visual instruction tuning (Liu et al., 2023), we introduce visual conditioning into the diffusion process. The overall architecture follows a widely adopted design comprising three components: a pretrained vision encoder, a lightweight vision-language connector, and a DLM decoder (LLaDA-8B).

Given an input image X_v and a corresponding textual instruction X_q, the objective is to generate the target text response X_a. First, we utilize a pretrained SigLIP-2 (Tschannen et al., 2025) as the vision encoder to extract visual features $Z_v = \Phi_v(X_v)$. A two-layer MLP with GELU activation serves as the connector Φ_c, projecting the visual features into the LLM's text embedding space to obtain continuous visual embeddings $H_v = \Phi_c(Z_v)$.

The visual embeddings H_v are concatenated with the discrete text embeddings of the instruction X_q and the response X_a to form the complete input sequence. During training, we apply the diffusion forward process *only* to the target response tokens X_a, leaving the image representations H_v and instruction tokens X_q uncorrupted as conditions. Our multimodal visual instruction tuning objective is thus formulated as:

$$\mathcal{L}_{\text{PerceptionDLM}_{\text{Base}}} = -\mathbb{E}_{(X_v,X_q,X_a),t,x_t}\left[\frac{1}{t}\sum_{i\in\mathcal{M}_a}\log p_\theta(x_0^i|x_t,H_v,X_q)\right], \tag{2}$$

where x_t is the partially masked target response sequence, and $\mathcal{M}_a$ denotes the indices of the masked tokens strictly within X_a.

PerceptionDLM-Base is trained using this discrete diffusion loss, followed by a 4-stage training paradigm. The detailed training parameters of each stage are shown in Table 3.

Dynamic Resolution for Multimodal Data To better handle high-resolution images and preserve fine-grained spatial details, we adopt a dynamic-resolution strategy following prior work (Chen et al., 2024b). Specifically, we dynamically partition each input image into a grid of 512×512 tiles based on the aspect ratio and resolution of the raw images. Optionally, if the number of tiles $n_{tiles} > 1$, we append an additional thumbnail of the original image to the end of the tiles list. Then each tile is independently processed with a pixel unshuffle operation to reduce the number of visual tokens to one-quarter of the original. Then the vision encoder encodes the preprocessed visual tokens and concatenates them into a sequence. This design lets the model scale to arbitrary image resolutions while balancing computational efficiency and spatial fidelity.

2.2 Training Strategies

We adopt a multi-stage training strategy for PerceptionDLM-Base, following the visual instruction tuning paradigm adapted for diffusion language models (You et al., 2025; Liu et al., 2023). Our training pipeline progressively improves multimodal alignment, general knowledge, and instruction-following ability in four stages. We use the latest open-source datasets to conduct our training.

Stage 1: Vision-Language Alignment. We first perform a lightweight alignment between visual and textual representations using the Bee-Training-Data-Stage1 (Zhang et al., 2025b). At this stage, we primarily train the vision-language connector while keeping most of the backbone parameters frozen, enabling stable alignment between visual features and the language embedding space.

Stage 2: Middle-stage Training. After alignment, we conduct large-scale training on Bee-Training-Data-Stage2 (Zhang et al., 2025b). We explore two training strategies at this stage and the subsequent stages: (1) full-parameter training, where both the diffusion backbone and vision encoder are updated, and (2) partially frozen training, where the vision encoder remains fixed while updating the language model and projection layers. More details can be found in Appendix B. This stage serves as a large-scale knowledge injection phase, exposing the model to diverse visual concepts and improving its general multimodal understanding capability. As in prior works (Li et al., 2025a; Zhang et al., 2025b), such large-scale mid-training helps bridge the gap between lightweight alignment and instruction tuning by enhancing representational capacity.

Stage 3: Instruction Tuning. We then perform supervised fine-tuning (SFT) using 22M samples from the LLaVA-OneVision-1.5-Instruct-Data (An et al., 2025), which contains diverse visual instruction tasks, including visual question answering, reasoning, OCR, and grounding. This stage enables the model to follow complex multimodal instructions and improves its generalization across a wide range of downstream tasks.

Stage 4: High-Quality SFT Refinement. Finally, we further refine the model using Honey-Data-15M (Zhang et al., 2025b), a high-quality multimodal instruction dataset enriched with dual-level chain-of-thought annotations. This dataset emphasizes complex reasoning and long-form responses, which significantly enhance the model's reasoning ability. This final stage improves both response quality and robustness in challenging multimodal scenarios.

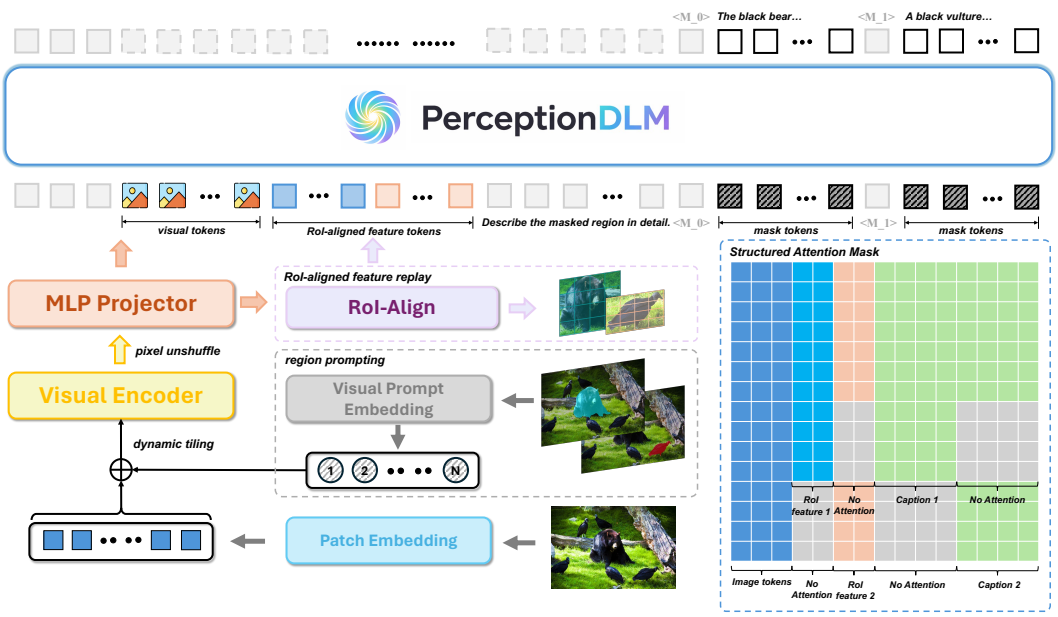


Figure 2: **Overview of the proposed parallel region perception architecture.** Built upon PerceptionDLM-Base, the model integrates region prompting, RoI-aligned feature replay, and a structured attention masking mechanism to enable parallel, disentangled caption generation for multiple regions within a single diffusion process.

3 Multimodal Diffusion Language Model for Parallel Perception

3.1 Task Formulation

Motivation. While recent MLLMs have achieved strong performance on visual understanding tasks, they typically rely on autoregressive decoding and generate outputs sequentially. This limitation becomes particularly evident in region-level perception tasks, where **multiple** regions need to be perceived and described. In contrast, diffusion language models naturally support parallel token generation, enabling joint prediction of multiple outputs within a **single denoising process**, whereas recent methods (Lian et al., 2025; Wang et al., 2025a) can only process each region independently. This motivates us to revisit region-level perception from the perspective of parallel generation.

Problem Definition. Given an image I and a set of N regions $\{R_i\}_{i=1}^N$, each represented by a binary mask, the goal is to generate a corresponding textual description $\{y_i\}_{i=1}^N$ for all regions. Existing approaches typically process each region independently: $y_i = f(I, R_i)$, which leads to linear growth in inference latency with respect to the number of regions (Rasheed et al., 2024; Lian et al., 2025). In contrast, leveraging the inherent parallel decoding ability of diffusion language models, we can model $\{y_i\}_{i=1}^N = f(I, \{R_i\}_{i=1}^N)$, enabling simultaneous captioning of all region descriptions within a single diffusion process.

3.2 Model Architecture

To enable efficient parallel region perception and captioning under the formulation above, we build upon the region-centric feature extraction paradigm of a recent strong AR-based baseline (Wang et al., 2025a). On top of this foundation, we propose a unified architecture that introduces region prompting and structured attention masking, as illustrated in Figure 2. Note that we highlight each component in a different color within the dashed boxes.

RoI-aligned Feature Replay from AR Baseline. For each region mask, localized visual features are extracted directly from the vision encoder and projected into the language embedding space as placeholder tokens. During preprocessing, each placeholder is expanded into a set of RoI feature tokens extracted from the corresponding visual region.

Region Prompting. While RoI features capture visual details, parallel generation requires the model to strictly distinguish multiple concurrent targets. To encode region identity effectively, we design a continuous region prompting mechanism. For each region R_i, we associate a learnable embedding e_i, which serves as a region-specific visual prompt. We spatially broadcast these embeddings and fuse them directly with the visual tokens corresponding to the masked regions, allowing the model to identify different regions and guide the generation process.

Structured Attention Masking Mechanism. Since DLMs denoise all tokens for all regions simultaneously, a key challenge in parallel generation is preventing interference across regions. To address this problem, we design a structured attention masking mechanism that enforces region-wise independence while preserving global context. Specifically, for tokens corresponding to region R_i, we restrict their attention to: (1) global visual tokens, (2) shared textual prompt tokens, (3) RoI feature tokens associated with region R_i, and (4) tokens within the same region-specific caption span. Meanwhile, attention to RoI features and caption tokens of other regions is masked out. This structured design results in a block-wise attention pattern that enforces region-level independence while preserving shared global context, enabling accurate parallel caption generation.

As illustrated in Figure 1, our model can simultaneously describe multiple masked regions while achieving substantially better perception efficiency. Overall, our architecture transforms region-level perception from a sequential process into a structured parallel generation problem, fully leveraging the intrinsic parallelism of diffusion language models.

3.3 PARADLC-BENCH

Inspired by DLC-Bench (Lian et al., 2025), we introduce an extended evaluation benchmark, **ParaDLC-Bench**, specifically adapted for multi-mask scenarios. Traditional evaluations rely on comprehensive reference captions, which struggle to handle the complex semantic entanglement that occurs during concurrent multi-target generation. Our benchmark inherits the reference-free evaluation paradigm of DLC-Bench, utilizing an LLM as a judge, but expands upon it: we extend the core focus of the evaluation from intra-region detail accuracy to inter-region feature independence and anti-interference capability. Similar to DLC-Bench, our evaluation process consists of two steps:

1. The model is prompted to generate detailed descriptions in parallel for multiple masked regions within a single image from the benchmark dataset.
2. An LLM serves as a judge, assessing the generated descriptions based on a predefined set of positive and negative questions. Furthermore, we observed that the reliability of the evaluation heavily depends on the reasoning capabilities of the judge LLM itself. Therefore, compared to DLC-Bench, which employs Llama-3.1-8B (Grattafiori et al., 2024), we utilize the more advanced GPT-5.2 (OpenAI, 2025) as our judge model to ensure accurate and robust assessments in complex multi-target interaction scenarios.

Regarding question design, we adopt the strategy from DLC-Bench and tailor it for parallel caption generation tasks:

- **Positive questions:** Focus on specific attributes unique to the target mask. The model earns a point for accurate inclusion, receives no points for omission, and incurs a penalty for factual errors.
- **Negative & Interference questions:** We introduce a key innovation within this category specifically for multi-target scenarios. In addition to checking for the inclusion of typical but absent attributes as in DLC-Bench, we specifically examine whether the model hallucinates features from other concurrent masked objects in the image into the current target's description (i.e., attribute entanglement). We award a point if the model avoids such cross-region hallucination; otherwise, it incurs a penalty. To prevent completely off-topic captions from receiving high scores, we award this point only if the model correctly recognizes the target object.

To ensure evaluation reliability, we implemented a rigorous quality control process. Expert human annotators conducted multiple rounds of strict cross-verification on all candidate questions to eliminate ambiguity, ensure strict spatial alignment with the targeted visual regions, and verify that the negative distractors were genuinely deceptive. We also verified the benchmark's robustness by re-evaluating

models across different open-source and closed-source judge LLMs (e.g., Qwen3.5-27B and Gemini-3.1-Pro), showing stable performance rankings (detailed in Appendix B). With this design, ParaDLC-Bench retains the flexibility and accuracy of DLC-Bench while effectively testing the model's true competence in handling dense, multi-target tasks. Our benchmark ultimately comprises 2345 manually verified questions, covering a wide range of multi-region attribute interactions and potential cases of hallucination. For more benchmark details, including the annotation pipeline and statistics, see Appendix D.

3.4 Training Data Engine

To enhance parallel-region captioning of DLMs, we further construct high-quality single-image, multi-mask caption data, which we refer to as **ParaCaption-5.7M**. We initially explore the Describe Anything dataset (Lian et al., 2025), which provides detailed caption annotations. However, because it lacks concurrent multi-mask samples, we design an automated data-construction pipeline based on existing large-scale segmentation datasets.

The training data comprises two different sources: the SA-1B dataset (Kirillov et al., 2023) and the COCONut dataset (Deng et al., 2024). For a selected subset of the massive SA-1B dataset, which contains many part-level regions unsuitable for general descriptions, we first filter out masks that are completely occluded (typically considered part-level). We then employ GAR-8B (Wang et al., 2025a), a state-of-the-art region captioning model, to generate initial descriptions and employ an LLM to extract the core categories. Next, using these extracted categories, we apply SAM3 (Carion et al., 2025) to re-predict the masks and discard low-matching samples by calculating the Intersection over Union (IoU) with the ground-truth masks. Meanwhile, for the COCONut dataset, which includes inherent mask and category annotations, we similarly use GAR-8B to generate initial descriptions and directly use Qwen3-8B (Yang et al., 2025a) to determine whether the generated text semantically matches the provided ground-truth category annotations. All verified data pairs from both datasets undergo a unified post-processing stage for length restriction and anti-repetition (hallucination) filtering to yield the final caption texts. Ultimately, we obtained 334k images with 3.4M masks from the COCONut dataset, and 83k images with 2.3M masks from the SA-1B dataset.

4 Experiment

4.1 Implementation Details of Parallel Region Perception Model

We adopt PerceptionDLM-Base as our base model for training a parallel region perception model, as it demonstrates strong perception capabilities across several frontier diffusion-based VLMs. During training, we set the maximum number of region prompts per image to 6 for training efficiency while ensuring the model learns multi-target perception. During evaluation, PerceptionDLM is under the default inference setting of 32 generation steps for a sequence length of 32 per mask. We use the same training loss as in Equation (2) with PerceptionDLM while setting all parameters trainable. We use transformers trainer and AdamW optimizer with a global batch size of 256 and a learning rate of 4×10^{-5}, with a linear warmup during the first 3% steps followed by a cosine decay schedule for the remaining steps. For RoI-aligned feature replay, we set the RoI output size to 4×4 in default.

4.2 Evaluation on Multimodal Benchmarks for PerceptionDLM-Base

As shown in Table 1, we evaluate PerceptionDLM-Base across 16 diverse multimodal benchmarks. Additional information and experiments are shown in Appendix B.

PerceptionDLM-Base consistently outperforms prior diffusion VLMs. PerceptionDLM-Base establishes a new, strong baseline among multimodal diffusion language models. Compared with frontier diffusion VLMs, our model delivers consistent and substantial gains on most benchmarks. Most notably, PerceptionDLM-Base outperforms LLaDA-V (You et al., 2025) on 15 out of the 16 evaluated benchmarks. Furthermore, when compared to more recent strong baselines such as SDAR-VL-8B (Cheng et al., 2025) and Dream-VL-7B (Ye et al., 2025), PerceptionDLM maintains a clear, comprehensive advantage, particularly in general VQA, fine-grained visual perception, and hallucination robustness. These gains indicate that the proposed training pipeline and multimodal architecture significantly enhance diffusion-based models' capabilities for general-purpose understanding and perception.

Competitive performance with advanced autoregressive VLMs. PerceptionDLM-Base is also competitive with recent autoregressive VLMs of similar size. Across the 16 benchmarks, PerceptionDLM-

Table 1: **Evaluation comparison across vision-language models on multimodal understanding benchmarks.** Most scores are reported as provided in the original papers. The symbol † denotes results we evaluated using official checkpoints and inference scripts, while ⋆ indicates results we evaluated from VLMEvalKit (Duan et al., 2024). "–" indicates that the corresponding metric is not reported in the original paper.

Benchmark	PerceptionDLM-Base 8B	LLaDA-V 8B	MMaDA 8B	LaViDa 8B	SDAR-VL 8B	Dream-VL 7B	Qwen2.5-VL 7B	InternVL3 8B
General VQA								
MMStar	**63.7**	60.1	–	–	59.9	59.9	63.9	68.2
SeedBench	**78.9**	74.8	64.2	–	75.5	76.4	77.0⋆	77.1⋆
MMBench	**85.0**	82.9	68.5	73.8	82.2	83.0	83.5	83.4
Reasoning								
MMMU	47.2	48.6	48.6	30.2	**53.0**	52.2	51.3⋆	57.3⋆
MathVista	**65.5**	52.4†	–	42.1	62.5	63.1	68.2	71.6
$\text{MathVerse}_{\text{Vision_Only}}$	25.3	20.6†	–	27.2	**36.5**	–	42.7⋆	23.1⋆
OCR & Doc & Chart								
AI2D	**85.0**	77.8	–	69.0	79.9	81.2	83.9	85.2
ChartQA	**91.6**	78.3	–	61.0	82.7	86.8	86.2	86.6
DocVQA	89.9	83.9	–	56.1	88.3	**94.4**	94.9	92.7
InfoVQA	74.6	66.3	–	36.2	73.2	**81.4**	82.6	76.8
Perception								
MMVP	**82.0**	76.7⋆	–	–	66.5	–	73.3⋆	80.0⋆
BLINK	**60.3**	50.9	–	–	–	52.9	55.3	55.5
RealWorldQA	**73.7**	63.2	–	–	66.5	66.3	68.4	70.8
CV-Bench-2D	**79.8**	77.7⋆	–	–	–	–	75.6⋆	79.0⋆
Others								
HallusionBench	**58.4**	50.9⋆	–	–	44.4	–	51.9	49.9
V⋆	**73.3**	62.8⋆	–	–	–	–	76.4⋆	67.5⋆

Table 2: **Comparison on the ParaDLC-Bench and DLC-Bench.** The symbol 1^* indicates that during the inference of these baseline diffusion VLMs, we set the number of denoising steps equal to the generation length, aiming to achieve their best possible generation quality.

Method	Size	ParaDLC-Bench					DLC-Bench		
		Pos (%)	Neg (%)	Avg (%)	TPF	Time (s)	Pos (%)	Neg (%)	Avg (%)
General VLMs									
GPT-5.2	–	38.0	71.0	55.2	–	–	18.0	60.8	39.4
Gemini-2.5-Pro	–	39.7	73.3	57.5	–	–	29.8	66.0	47.9
Gemini-3.1-Pro	–	43.6	81.1	63.7	–	–	26.2	65.8	46.0
AR-based Region-specific VLMs									
PixelRefer	7B	40.8	78.7	60.5	1	718	49.6	87.0	68.3
DAM	3B	48.1	87.2	69.2	1	326	52.3	82.2	67.3
GAR	8B	49.0	87.6	69.5	1	479	50.9	84.6	67.8
Diffusion-based VLMs									
LLaDA-V	8B	24.1	46.3	35.2	1^*	3241	10.0	39.2	24.6
SDAR-VL	8B	30.2	28.8	31.3	1^*	945	9.9	47.6	28.8
Dream-VL	7B	29.7	28.6	30.4	1^*	446	9.8	39.6	24.7
Ours Parallel Caption Model									
PerceptionDLM	8B	42.3	82.4	62.4	2.9	276	33.4	72.8	53.1

Base achieves superior or comparable scores on a majority of the tasks. It excels particularly in fine-grained visual perception, systematically outperforming both Qwen2.5-VL-7B (Bai et al., 2025b) and InternVL3-8B (Zhu et al., 2025) in these specific areas. This suggests that PerceptionDLM-Base possesses a distinct advantage in region-sensitive and detail-oriented visual understanding. While a performance gap remains in complex, reasoning-heavy scenarios (e.g., MMMU (Yue et al., 2024) and

MathVista (Lu et al., 2024)), we observe that arbitrary-order parallel decoding fundamentally limits the reasoning potential of diffusion language models (Ni et al., 2026). Thus, we adopt autoregressive-order decoding for PerceptionDLM-Base during these mathematical reasoning evaluations to better preserve reasoning traces. Inspired by recent advancements like DeepSeek-R1 (Guo et al., 2025), this bottleneck highlights a clear avenue for future work: leveraging Reinforcement Learning (RL) to further unlock the reasoning potential of diffusion-based VLMs.

4.3 Evaluation on Captioning Benchmarks with PerceptionDLM

To further assess fine-grained multimodal perception, we evaluate PerceptionDLM on region captioning benchmarks, including ParaDLC-Bench and the single-region DLC-Bench (Lian et al., 2025).

Superior region captioning capability among diffusion VLMs. As shown in Table 2, PerceptionDLM exhibits leading advantages over existing diffusion-based VLMs. On ParaDLC-Bench, it achieves an average accuracy of 62.4%, nearly doubling the performance of SDAR-VL (Cheng et al., 2025) (31.3%) and LLaDA-V (You et al., 2025) (35.2%). This substantial margin is consistent on DLC-Bench (Lian et al., 2025) (53.1% vs. 24.6% for baselines).

Competitive performance with unprecedented efficiency. To analyze efficiency, we adopt the Tokens Per Forward (TPF) metric (Qian et al., 2026), which quantifies the average number of generated tokens per forward pass. Compared with AR-based region-specific models (e.g., DAM and GAR), PerceptionDLM demonstrates highly competitive accuracy while unlocking massive speed advantages. Although its average accuracy on ParaDLC-Bench is slightly lower than AR-based models, PerceptionDLM drastically reduces the total inference time across the benchmark to just 276 seconds, compared to 479 seconds for GAR and 718 seconds for PixelRefer (Yuan et al., 2025c).

This efficiency is driven by our parallel decoding paradigm. While AR models and standard DVLM baselines are restricted to a TPF of 1, PerceptionDLM achieves a TPF of 2.9, generating multiple region captions simultaneously. Note that on DLC-Bench, where instances contain only a single mask, our parallel-processing advantage cannot be fully exploited. Nevertheless, PerceptionDLM still sets a new benchmark for diffusion-based VLMs across both settings.

To systematically analyze the computational advantages of PerceptionDLM, we conduct a speed and efficiency profiling. As shown in Figure 1(b), PerceptionDLM achieves near-linear TPS growth while maintaining stable per-image latency ($\sim$2.9s). In contrast, GAR-8B is bottlenecked at a nearly constant TPS, and its latency degrades approximately linearly with the number of regions. As shown in Figure 1(c), under a constant heavy workload (4 masks per image), increasing the degree of parallelism (masks processed per pass) yields strong scaling for PerceptionDLM: throughput improves by 3.44 times, and single-image latency drops from 10.04s to 2.92s when fully parallelized.

4.4 Step Distillation for Inference Acceleration

PerceptionDLM parallelizes region captions but still requires multiple denoising passes. We therefore apply pseudo-trajectory distillation following d3LLM (Qian et al., 2026) to PerceptionDLM-Base and PerceptionDLM. Here, the pseudo-trajectory refers to the token decoding order extracted from the teacher's generation, while ground-truth responses provide the training target. Detailed methodology, training settings and complete results are provided in Appendix C.

As shown in Tables 13 and 14, step distillation enables faster decoding for both models, with a trade-off in generation quality. For PerceptionDLM-Base, the distilled model with entropy-based multi-block decoding achieves a $4.6\times$ wall-clock throughput speedup, with a limited reduction in benchmark performance. For PerceptionDLM, combining distillation with entropy-threshold decoding yields up to a $1.9\times$ TPF speedup with only a small reduction in caption quality.

5 Conclusion

We present PerceptionDLM, a diffusion-based multimodal model for parallel region perception. Built on a stronger diffusion VLM baseline, it generates multiple region captions in a single denoising process rather than decoding them one by one. This design preserves competitive caption quality while substantially improving multi-region perception efficiency, which further compounds with step-level distillation. We also introduce ParaDLC-Bench to evaluate both caption accuracy and inference speed in parallel localized captioning. Overall, we hope our results establish diffusion-based multimodal models as a promising direction for efficient fine-grained visual understanding and a new way of caption generation.

AI USE STATEMENT

In this work, large language models were used as components of the automated data-construction, attribute-extraction, and benchmark-evaluation pipelines described in the paper. AI-assisted tools were also used for proofreading, grammatical correction, language refinement, and code implementation. The authors designed the research questions, methods, experiments, and scientific claims, reviewed all AI-assisted outputs, and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper specifies the model architecture, training objective, datasets, evaluation protocols, and principal implementation settings. Additional training details, ablations, benchmark construction procedures, prompts, and qualitative results are provided in the appendix. Code, model checkpoints, and data resources will be released to support reproduction.

REFERENCES

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, et al. SDAR: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*, 2025.

Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *CVPR*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMevalKit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, 2024.

Chaoyou Fu et al. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

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

Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. Dataseg: Taming a universal multi-dataset multi-task segmentation model. *NeurIPS*, 36:67329–67354, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Aniruddha Kembhavi, Michael Salvato, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. Ai2d: A dataset for diagram understanding. In *CVPR*, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *ICCV*, 2025.

Bo Li, Peiyuan Li, Zhaolin Zhang, Yifan Wang, Yinan Wang, Zhengyuan Liu, Kai Chen, and Ziwei Liu. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2025a.

Xiangtai Li, Tao Zhang, Yanwei Li, Haobo Yuan, Shihao Chen, Yikang Zhou, Jiahao Meng, Yueyi Sun, Shilin Xu, Lu Qi, et al. Denseworld-1m: Towards detailed dense grounded caption in the real world. *arXiv preprint arXiv:2506.24102*, 2025b.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *ICCV*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

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? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

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

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021a.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Infographicvqa. *arXiv preprint arXiv:2104.12756*, 2021b.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, et al. The flexibility trap: Why arbitrary order limits reasoning potential in diffusion language models. In *ICML*, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *NeurIPS*, 2025.

OpenAI. Openai-gpt-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025.

Yu-Yang Qian, Junda Su, Lanxiang Hu, Peiyuan Zhang, Zhijie Deng, Peng Zhao, and Hao Zhang. d3LLM: Ultra-fast diffusion llm using pseudo-trajectory distillation. *arXiv preprint arXiv:2601.07568*, 2026.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pp. 8430–8439, 2019.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *NeurIPS*, 2024a.

Shengbang Tong et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024b.

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

Haochen Wang, Yuhao Wang, Tao Zhang, Yikang Zhou, Yanwei Li, Jiacong Wang, Jiani Zheng, Ye Tian, Jiahao Meng, Zilong Huang, et al. Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. *arXiv preprint arXiv:2510.18876*, 2025a.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *ICCV*, 2025b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. In *NeurIPS*, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025a.

Haobo Yuan, Yueyi Sun, Yanwei Li, Tao Zhang, Xueqing Deng, Henghui Ding, Lu Qi, Anran Wang, Xiangtai Li, and Ming-Hsuan Yang. Visual reasoning tracer: Object-level grounded reasoning benchmark. *arXiv preprint arXiv:2512.05091*, 2025b.

Yuqian Yuan, Wenqiao Zhang, Xin Li, Shihao Wang, Kehan Li, Wentong Li, Jun Xiao, Lei Zhang, and Beng Chin Ooi. Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity. *arXiv preprint arXiv:2510.23603*, 2025c.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Renrui Zhang et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, 2024a.

S Zhang, P Sun, S Chen, M Xiao, W Shao, W Zhang, Y Liu, K Chen, and P Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *NeurIPS*, 2024b.

Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. *arXiv preprint arXiv:2504.10465*, 2025a.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. *arXiv preprint arXiv:2510.13795*, 2025b.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, et al. Samtok: Representing any mask with two words. *arXiv preprint arXiv:2601.16093*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Overview

This appendix is organized as follows:

- In **Appendix B**, we provide more implementation details along with additional experimental results. Detailed ablation studies analyzing the impact of each core component of PerceptionDLM can also be found in this section.
- In **Appendix C**, we study step distillation of PerceptionDLM, transferring pseudo-trajectory distillation to both PerceptionDLM-Base and the region-captioning model, and analyzing the resulting quality–speed trade-off under multi-block decoding.
- In **Appendix D**, we introduce the details of our proposed ParaDLC-Bench, including the specific annotation guidelines, quality control protocols, and overall dataset statistics.
- In **Appendix E**, we provide an extended discussion of related works, covering broader literature on diffusion language models and region-level multimodal perception and captioning.
- In **Appendix F**, we provide extensive visualization and qualitative results, showcasing PerceptionDLM's performance on fine-grained image understanding and parallel region captioning tasks.
- In **Appendix G**, we discuss the potential limitations of our current approach and provide an analysis of typical failure cases.

B Additional Experiments

Implementation details. We summarize the implementation details of PerceptionDLM-Base pretraining and the parallel captioning model in this section.

The four-stage training setup of PerceptionDLM-Base is listed in Table 3; the entire pipeline is trained on 32×NVIDIA H100 (80GB) GPUs and takes approximately three weeks in total. We use the AdamW optimizer with cosine learning-rate schedule and 3% linear warmup, BF16 mixed precision, and gradient checkpointing. The vision encoder is SigLIP-2 (Tschannen et al., 2025) and the diffusion backbone is initialized from LLaDA-Instruct-8B (Nie et al., 2025); images are processed with the dynamic-resolution tiling strategy using a tile size of 512×512.

For the parallel captioning model, we initialize from PerceptionDLM-Base and set all parameters trainable. We use a global batch size of 256, learning rate of 4×10^{-5}, one training epoch, maximum sequence length of 4096, and RoI output size of 4×4. We set the maximum number of region prompts per image at 6 during training. Training on the full ParaCaption-5.7M corpus takes about 2 days on 32×H100 GPUs and since it starts from the strong base checkpoint, this stage is substantially cheaper than base pretraining.

For parallel-captioning ablations, unless otherwise specified, all runs use the same setting: training on DAM (Lian et al., 2025) only, default PerceptionDLM as baseline architecture, and inference with 32 diffusion steps per diffusion process and generation length 32 per mask. All latency and throughput evaluations are conducted on a single H100 GPU using BF16 precision, with a strictly matched batch size of 1 per image for all models. We keep all other hyper-parameters identical across ablation runs to ensure fair comparison. For evaluation protocols, multimodal benchmarks are run with VLMEvalKit (Duan et al., 2024). For ChartQA, DocVQA, and InfoVQA, we additionally adopt a stricter answer-matching judge based on a locally deployed Qwen3-8B (Yang et al., 2025a) to better reflect the true accuracy. For region-captioning benchmarks, the LLM judge (GPT-5.2 by default) is queried at `temperature=0` to remove sampling noise, and we further report robustness across different judge models in Table 8.

Evaluation on Multimodal Benchmarks We evaluate PerceptionDLM-Base on a diverse set of multimodal benchmarks covering general VQA, reasoning, chart and document understanding, perception-oriented tasks, and hallucination robustness. Specifically, we report results on **General VQA** benchmarks (MMStar (Chen et al., 2024a), SeedBench (Li et al., 2024), MMBench (Liu et al., 2024)), **Reasoning** benchmarks (MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), MathVerse-Vision-Only (Zhang et al., 2024a)), **OCR, Chart and Document Understanding** benchmarks (AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021a), InfoVQA (Mathew et al., 2021b)), **Perception** benchmarks (MMVP (Tong et al., 2024b),

BLINK (Fu et al., 2024), RealWorldQA (xAI, 2024) , CV-Bench-2D (Tong et al., 2024a)), and **Others** (HallusionBench (Guan et al., 2024), V* (Wu & Xie, 2024)).

Table 3: Training parameters of PerceptionDLM-Base

	Stage-1	Stage-2	Stage-3	Stage-4
Dataset #Samples	Bee-Training-Data-Stage1 1M	Bee-Training-Data-Stage2 14M	LLaVA-OneVision-1.5-Instruct-Data 22M	Honey-Data-15M 15M
Vision Tower LLM Backbone Trainable Model Parameters	siglip2-so400m-patch16-512 LLaDA-Instruct-8B Projector	siglip2-so400m-patch16-512 LLaDA-Instruct-8B Projector + LLM backbone	siglip2-so400m-patch16-512 LLaDA-Instruct-8B Projector + LLM backbone	siglip2-so400m-patch16-512 LLaDA-Instruct-8B Projector + LLM backbone
Batch Size Model Max Length Learning Rate Epoch	512 2048 1×10^{-3} 1	512 4096 4×10^{-5} 1	512 4096 4×10^{-5} 1	512 4096 4×10^{-5} 1

Table 4: Multimodal benchmark comparison between PerceptionDLM-Base with trainable ViT variant and the frozen-ViT baseline. The compared benchmarks follow the same protocol as in Table 1.

Model Variant	**MMStar**	**MMBench**	**MMMU**	**MathVista**	**AI2D**	**ChartQA**	**MMVP**	**HallusionBench**
PerceptionDLM-Base (Train ViT)	61.9	83.7	**48.7**	61.9	82.0	89.3	79.7	54.6
PerceptionDLM-Base (Freeze ViT, baseline)	**63.7**	**85.0**	47.2	**65.5**	**85.0**	**91.6**	**82.0**	**58.4**

Impact of ViT Training Strategy on PerceptionDLM-Base. We explore two optimization strategies for instruction tuning: (1) full-parameter training (updating both the diffusion backbone and ViT), and (2) partially frozen training (freezing ViT). Table 4 evaluates their impact on multimodal capabilities. The frozen-ViT baseline significantly outperforms the full-parameter variant across most benchmarks. Thus, we adopt the partially frozen strategy to preserve broad visual understanding.

Table 5: Data scaling ablations for Parallel Captioning. Increasing the amount and diversity of training data consistently improves performance.

Method	**Training Data**	**Pos(%)** ↑	**Neg(%)** ↑	**Avg(%)** ↑
PerceptionDLM	DAM	33.8	73.6	53.7
	DAM + COCONut	38.4	77.1	57.7
	DAM + COCONut + SAM	**42.3**	**82.4**	**62.4**

Ablations on Data Scaling Table 5 demonstrates consistent performance gains as the training data scales up. Starting with the DAM dataset baseline (53.7% average accuracy), we observe that introducing processed data sourced from the COCONut dataset (Deng et al., 2024) improves the average accuracy to 57.7%. Further expanding the training corpus with SA-1B(SAM) dataset (Carion et al., 2025) annotations yields the best performance, reaching 62.4%. Notably, the improvements indicate that both data scale and data diversity are critical.

Ablations on Vision Encoder Training Strategy. We investigate whether unfreezing the vision encoder during visual instruction tuning benefits region-level perception. As shown in Table 6, fully updating the vision encoder results in performance degradation on the fine-grained ParaDLC-Bench, consistent with observations on multimodal benchmarks. This suggests that the robust representations learned during the vision encoder's large-scale pre-training are highly important for complex localized perception tasks.

Ablations on Architecture Designs. In Table 7, we evaluate the necessity of our proposed structural designs for parallel captioning under a controlled training setting. The results confirm that all three core modules are indispensable:

- Region Prompting: Without explicit region prompting, the model loses its spatial grounding capability entirely, resulting in a catastrophic drop to 1.1% average accuracy. The model fails to associate textual generation with specific visual targets.
- Structured Attention: Replacing our structured attention with standard full attention degrades the average accuracy by 6.2%. This validates that structured attention successfully isolates

Table 6: Ablations on the training strategy of the vision encoder in PerceptionDLM. Freezing the vision encoder performs better, and is thus used as the default baseline setting.

Method	Vision Encoder	Pos(%) ↑	Neg(%) ↑	Avg(%) ↑
PerceptionDLM	Train	31.8	68.8	50.3
	Freeze	**33.8**	**73.6**	**53.7**

Table 7: Ablations on core design choices for parallel captioning under the same training setting, including module design and attention masking strategy. Results show that *region prompting*, *RoI-aligned feature replay*, and *structured attention* are all important for the final performance. Following the judge-scoring protocol, Pos/Neg scores can be slightly negative when penalties dominate.

Method	Model Setting	Pos.(%) ↑	Neg.(%) ↑	Avg.(%) ↑
PerceptionDLM	w/o region prompting	-0.01	2.4	1.1
	w/o RoI-aligned feature replay	29.9	72.7	51.3
	Full Attention	30.8	64.1	47.5
	Full model (default baseline)	**33.8**	**73.6**	**53.7**

the processing of different masks during parallel generation, mitigating feature interference and target confusion across multiple concurrent descriptions.

- RoI-aligned Feature Replay: Removing this module results in a 2.4% accuracy drop. This indicates that explicitly pooling and aligning localized visual features provides fine-grained cues that are essential for accurate regional perception.

Qualitative Breakdown of Region Prompting. Removing region prompting causes a catastrophic grounding failure, though not a language breakdown. Below is a qualitative example from the image in Figure 1(a) without learnable region prompts:

- **Mask 1:** "A tall, cylindrical stone pillar with a smooth surface and a slightly tapered top. The pillar has a uniform texture and appears to be made of a solid material."
- **Mask 2:** "A tall, cylindrical stone pillar with a smooth surface and a slightly tapered top..."
- **Mask 3:** "A tall, cylindrical stone pillar with a smooth surface and a slightly tapered top..."

The model emits grammatically fluent, identical text for all masks. Without learnable region prompts, the model loses the ability to bind specific generated tokens to distinct spatial coordinates, completely failing the localized perception task.

Judge Model Sensitivity on ParaDLC-Bench. To establish reproducibility and test the robustness of our benchmark, we re-evaluate the models using different Judge LLMs, including the open-source Qwen3.5-27B and the closed-source Gemini-3.1-Pro. As shown in Table 8, although there are minor fluctuations in absolute scores and partial rankings due to differences in the judge models' reasoning capabilities, the main conclusions remain unchanged. PerceptionDLM consistently outperforms all other diffusion Vision-Language Models, remaining highly competitive with traditional AR models. This fully validates the reliability of the ParaDLC-Bench evaluation framework.

Impact of Denoising Steps on Accuracy and Latency. To illustrate the trade-off between accuracy and inference latency, we evaluate PerceptionDLM on ParaDLC-Bench using Qwen3.5-27B judge across varying diffusion steps while keeping the generation length fixed at 32. As shown in Table 9, `steps=32` is the optimal operating point, offering the best balance of high captioning accuracy and efficient inference time.

Inference with Exceeded Visual Prompts. To address how the model generalizes to mask counts exceeding the number of trained visual prompts, we conduct an additional experiment testing a forced edge case where the model reuses the exact same visual prompt for all masks during inference. As shown in Table 10, forcing the model to reuse visual prompts for distinct targets results in a performance drop. However, this is a graceful degradation (the average accuracy evaluated by

Table 8: Performance on ParaDLC-Bench by different judge models (G3.1P = Gemini-3.1-Pro Judge; Q3.5 = Qwen3.5-27B Judge).

Model	G3.1P-Pos(%)	G3.1P-Neg(%)	G3.1P-Avg(%)	Q3.5-Pos(%)	Q3.5-Neg(%)	Q3.5-Avg(%)
GPT-5.2	42.6	72.6	57.6	50.7	72.8	61.8
Gemini-2.5-Pro	41.1	73.6	57.3	45.5	73.4	59.4
Gemini-3.1-pro	45.3	82.7	64.0	53.0	82.8	67.9
PixelRefer	46.9	87.7	67.3	57.3	88.8	73.1
DAM	52.3	88.0	70.1	59.6	88.8	74.2
GAR	53.5	88.8	71.2	59.7	89.7	74.7
LLaDA-V	31.0	53.0	42.0	38.9	51.0	45.0
SDAR-VL	37.8	32.7	35.3	44.4	38.1	41.3
Dream-VL	35.4	35.8	35.6	41.4	37.9	39.7
PerceptionDLM	47.1	84.9	66.0	59.4	87.5	73.5

Table 9: Impact of denoising steps on accuracy and latency on ParaDLC-Bench.

Method	Steps	Pos(%) ↑	Neg(%) ↑	Avg(%) ↑	Time (s)
PerceptionDLM	16	54.5	83.4	69.0	138
	32	**59.4**	**87.5**	**73.5**	**276**
	48	57.3	83.7	70.5	411
	64	51.2	83.8	67.5	549

Qwen3.5-27B drops from 73.5% to 68.6%), indicating that the model retains a strong baseline of parallel perception even when prompt capacity is exceeded.

For practical deployment in highly dense scenarios, the recommended strategy is to split the masks into multiple inference passes. As demonstrated by the parallelism scaling in Figure 1(c), this chunking strategy still yields a significant speedup over sequential AR pipelines while bypassing potential long-context limitations. Naturally, if a specific downstream application strictly requires simultaneous dense perception, the visual prompt capacity can easily be scaled up during training to accommodate it.

Table 10: Inference with Exceeded Visual Prompts (Evaluated using Qwen3.5-27B Judge).

Model Setting	Q3.5-Pos(%) ↑	Q3.5-Neg(%) ↑	Q3.5-Avg(%) ↑
PerceptionDLM (same visual prompt)	52.8	84.4	68.6
PerceptionDLM (baseline)	**59.4**	**87.5**	**73.5**

Ablations on Caption Length Scaling. To investigate the impact of description length, we conduct an ablation study by scaling the maximum number of generation tokens per mask from 32 to 64, while keeping the diffusion steps fixed at 32. As shown in Table 11, with fixed denoising steps, generating longer sequences forces the model to decode more tokens per step. This increased decoding density leads to severe error accumulation along the denoising trajectory, resulting in semantic drifting.

Single-Region Quality Regression Ablation. To investigate whether parallel training degrades single-region capabilities, we conduct an ablation study training PerceptionDLM exclusively on single-mask format data. As shown in Table 12, single-mask training yields performance similar to our parallel training baseline. This indicates that the gap compared to AR models on single-region tasks is not a regression caused by parallelization.

Zero-Shot Performance of PerceptionDLM-Base on ParaDLC-Bench. To disentangle the benefits of our proposed architectural design from the supervisory signals provided by GAR-8B in the ParaCaption-5.7M dataset, we evaluate the zero-shot region perception capability of PerceptionDLM-Base prior to any dataset-specific fine-tuning. Evaluated by the GPT-5.2 judge, PerceptionDLM-Base achieves a zero-shot average accuracy of **53.0%** on ParaDLC-Bench. This baseline substantially outperforms existing DVLMs, such as SDAR-VL, which attains an average accuracy of 31.3%(shown

Table 11: Caption length scaling on Parallel Captioning(Evaluated using Qwen3.5-27B Judge).

Method	Length	Q3.5-Pos(%) ↑	Q3.5-Neg(%) ↑	Q3.5-Avg(%) ↑
PerceptionDLM	64	55.5	80.4	68.0
	32	**59.4**	**87.5**	**73.5**

Table 12: Single Mask Training Ablation on DLC-Bench.

Method	Pos(%) ↑	Neg(%) ↑	Avg(%) ↑
PerceptionDLM (single mask training)	**34.5**	69.8	52.1
PerceptionDLM (baseline)	33.4	**72.8**	**53.1**

in Table 2). This demonstrates that the significant performance improvements of our framework are fundamentally driven by the architectural advancements, rather than merely stemming from the high-quality synthetic training corpus.

C STEP DISTILLATION OF PERCEPTIONDLM

PerceptionDLM eliminates the sequential cost of describing regions individually, but every denoising step is still a full forward pass over the multimodal sequence. Region-level parallelism and step-level parallelism are therefore orthogonal, and the latter becomes the dominant cost once responses grow long. d3LLM (Qian et al., 2026) reduces the number of denoising steps by distilling the decoding order of a teacher model. We transfer this recipe to the multimodal setting, applying it to PerceptionDLM-Base for general multimodal understanding and to PerceptionDLM for parallel region captioning.

Method. Given a prompt $\mathbf{x}$ and a maximum output length n, a teacher dLLM generates a trajectory $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ by accepting one token at each step. We use this trajectory to determine the token decoding order and take the training targets from the ground-truth response $\mathbf{y}$, following d3LLM (Qian et al., 2026). For a mask ratio $t \in [0, 1]$ and a decoding window $w = \{s, \ldots, s + k\}$, the noisy training sequence $\widetilde{\mathbf{y}} \in \mathbb{R}^n$ is constructed as

$$[\widetilde{\mathbf{y}}]_i = \begin{cases} [\mathbf{y}]_i, & \text{if } i \le s \text{ or } \left[\mathcal{T}_{s+\lceil kt \rceil}\right]_i \neq \texttt{mask}, \\ \texttt{mask}, & \text{if } i > s + k \text{ or } \left[\mathcal{T}_{s+\lceil kt \rceil}\right]_i = \texttt{mask}. \end{cases} \tag{3}$$

Here, `mask` is the mask-token ID and $[\cdot]_i$ denotes the token at position i. At a sampled denoising step, decoded positions are filled with ground-truth tokens, while the remaining positions stay masked. The teacher trajectory provides the decoding order, and $\mathbf{y}$ provides the training targets. The student learns this order by minimizing the cross-entropy loss on the masked tokens. During training, we gradually increase the mask ratio from 0 to 0.8 and enlarge the decoding window. We train on both the sampled masking pattern and its complement, with entropy regularization using a temperature of 0.5 and a weight of 2.0.

For multimodal training, pseudo-trajectory masking is applied only to the answer tokens X_a, while the visual embeddings H_v and instruction tokens X_q remain uncorrupted conditions, as in Equation (2). At inference, we apply entropy-based multi-block decoding over response blocks for PerceptionDLM-Base and region-caption spans for PerceptionDLM.

Accelerated Multimodal Decoding via PerceptionDLM-Base. We generate teacher trajectories with PerceptionDLM-Base on 300K samples from Honey-Data-15M. Trajectory generation uses 128 diffusion steps, with an output length of 128 tokens and a block length of 32. We then update the language model for one epoch at a learning rate of 1×10^{-5}, keeping the vision encoder and visual abstractor frozen, while enlarging the training window from 16 to 32 tokens.

Table 13 presents the general multimodal results for the distilled model at τ=0.9. At τ=0.9, entropy-based multi-block decoding increases the wall-clock throughput of the distilled PerceptionDLM-Base

Table 13: **Step distillation on PerceptionDLM-Base.** "+ Distill" denotes pseudo-trajectory distillation and "mb" denotes entropy-based multi-block decoding at threshold τ. TPF and TPS are measured on long-form VQA with a generation length of 128.

Model	MMStar	MMBench	MathVista	AI2D	ChartQA	DocVQA	MMVP	RealWorldQA	TPF	TPS
PerceptionDLM-Base	63.7	85.0	65.5	85.0	91.6	89.9	82.0	73.7	1.00	9.6
+ Distill (mb τ=0.9)	61.0	81.7	60.3	82.7	80.8	84.3	79.7	72.8	**3.11**	**43.9**

by $4.6\times$. Its scores remain close to those of native decoding on most benchmarks, suggesting that accelerated decoding introduces limited additional degradation in these cases.

Further Efficiency Gains in Parallel Region Captioning. We sample 300K images from ParaCaption-5.7M and use the original PerceptionDLM to generate teacher trajectories in the same parallel-region format. During trajectory generation, each region caption has a maximum length of 64 tokens and uses a block length of 32 tokens. We then train a LoRA adapter (r=64, α=128) for one epoch, with the block size annealed from 8 to 32. At inference, we apply entropy-threshold decoding to all region captions in parallel. Each forward pass fills masked positions with predictive entropy below τ, selecting the lowest-entropy position in each unfinished region if none meets the threshold. Decoding terminates when no active masked positions remain.

Table 14: **Step distillation on PerceptionDLM** for parallel region captioning on ParaDLC-Bench. "Native (32)" uses 32 denoising steps. τ is the predictive-entropy threshold used to decode each 32-token region caption; a larger value accepts more tokens per forward pass. NFE denotes the total number of model forward passes, and speedup is the TPF ratio relative to the original model under native decoding.

Model	Decoding	Pos(%) ↑	Neg(%) ↑	Avg(%) ↑	NFE	TPF	Time (s)	Speedup (TPF)
PerceptionDLM	Native 32	42.3	82.4	62.4	3296	2.90	276	1.00×
PerceptionDLM	τ=0.5	39.0	79.7	59.3	2844	3.36	242	1.16×
+ Distill	τ=0.5	40.9	79.8	60.4	2240	4.27	191	1.47×
+ Distill	τ=0.9	38.6	81.1	59.8	**1713**	**5.59**	**149**	**1.92**×

As shown in Table 14, we evaluate the distilled PerceptionDLM under entropy-threshold decoding with variable entropy thresholds. Combining step distillation with entropy-threshold decoding substantially improves the inference efficiency of parallel region captioning, with a small reduction in quality. A moderate threshold provides a favorable balance between accuracy and efficiency, while a more aggressive threshold approaches a $2\times$ TPF speedup at a larger quality cost.

D Details of ParaDLC-Bench

D.1 Image and Instance Selection

To ensure the diversity and complexity of the evaluation scenarios, our base images are sourced from two mainstream open-source datasets: the validation set of Objects365 V2 (Shao et al., 2019) and the DaTaSeg Objects365 Instance Segmentation Dataset (Gu et al., 2023). Both datasets provide a large number of images with high-quality instance segmentation masks. Specifically, the Objects365 V2 subset includes 54 images with a total of 178 mask instances, while the DaTaSeg subset comprises 46 images with a total of 121 mask instances.

Regarding the instance sampling and filtering criteria, we primarily rely on the professional judgment of human annotators. Unlike DLC-Bench (Lian et al., 2025), which focuses on a single challenging target, ParaDLC-Bench specifically emphasizes the interaction potential among multiple targets. Annotators consciously select "multi-instance combinations" that are spatially adjacent, semantically confusing, or prone to feature entanglement. Finally, to ensure the benchmark's rigor, we performed strict data deduplication and isolation, ensuring that none of the images used in this benchmark were present in the training sets of the evaluated models, thereby preventing data leakage.

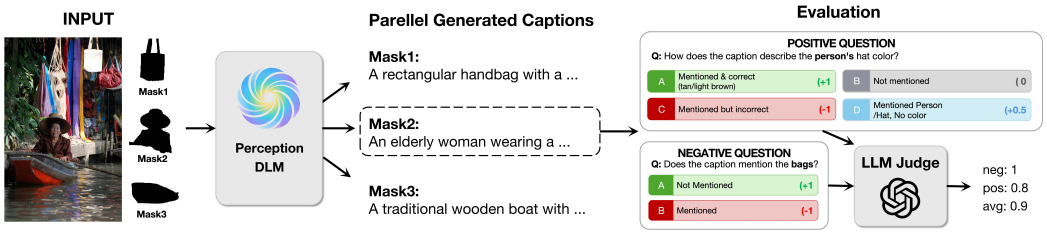


Figure 3: **The evaluation pipeline for detailed localized captioning in ParaDLC-Bench.** (a) During evaluation, given an input image and multiple region masks, the Perception DLM simultaneously generates textual descriptions for all specified regions in a single pass. (b) Subsequently, each generated caption, paired with its corresponding evaluation questions, is fed into a text-only large language model (LLM Judge) to calculate the final score. (c) In the scoring mechanism, positive questions reward accurate details and penalize factual errors, while negative questions strictly penalize mislocalization and hallucinations to prevent false positives in irrelevant areas.

D.2 Data Annotation Pipeline

The question generation pipeline of ParaDLC-Bench is inspired by the methodology of DLC-Bench (Lian et al., 2025), where the goals are designed specifically for "multi-target scenarios". The entire process consists of three core components: "LLM Extraction - Human Curation - Automated Format Conversion". We uniformly use the more advanced GPT-5.2 (OpenAI, 2025) as the foundational model for complex logical reasoning. We provide all of our prompts utilized in constructing our ParaDLC-Bench in Figures 8 to 14.

Attribute Extraction: We input the cropped version of the original image alongside the segmented image (retaining only the target instance) into GPT-5.2. For positive attribute extraction, the model is required to extract all visible parts of the target and generate an attribute list covering dimensions such as color, shape, texture, material, and size, uniformly outputting a tuple in the format of `([object name], [part name], [property name], [property value])`. Regarding negative attribute extraction—our core innovative design—we construct challenging distractors tailored for multi-target scenarios by extracting two types of negative samples that easily trigger model hallucinations: first, *Negative Parts*, which are parts commonly associated with the object but invisible in the current mask; second, *Salient Negatives / Mislocalization Targets*. In ParaDLC-Bench, the latter not only includes standard background objects but also focuses heavily on other concurrently evaluated targets in the image. This aims to test whether the model will incorrectly assign the features of other concurrent targets to the current one.

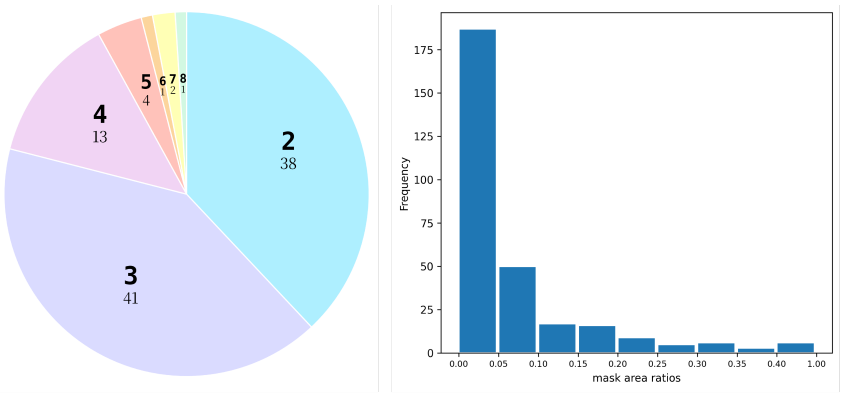
Human Curation & Refinement: To compensate for factual errors made by the LLM, human experts perform strict cross-verification starting from the initial lists generated by GPT-5.2. Experts are responsible for supplementing missing salient attributes and removing ambiguous details, ensuring that all attributes strictly align with the masked region and that the distractors are highly deceptive.

Question Generation: After obtaining the manually verified attribute lists, we utilize GPT-5.2 to convert them into mutually exclusive multiple-choice questions. Through this rigorous pipeline, we ultimately constructed an evaluation question bank comprising 2,345 high-quality multiple-choice questions.

D.3 Scoring Mechanism

During the evaluation phase, we inherit the reference-free judge evaluation paradigm from DLC-Bench (Lian et al., 2025). The evaluated models are required to generate descriptions for multiple specified masked regions in the image. For models equipped with parallel generation capabilities, descriptions for all regions can be generated concurrently in a single pass; for models lacking this feature, descriptions are generated sequentially for each masked region. Subsequently, an LLM Judger (GPT-5.2 (OpenAI, 2025)) evaluates these descriptions using predefined positive and negative questions. Regarding judge variance, GPT-5.2 evaluations are conducted at strict `temperature=0` to eliminate sampling noise.

To prevent models from cheating to gain higher scores by generating excessively long and ambiguous descriptions, a model can only receive positive or negative points if the base target is correctly identified. Unlike DLC-Bench, which averages global positive and negative scores, we compute the final score by averaging the individual scores across all masks, ensuring equal evaluation weighting for each region. The overall scoring process and an example are illustrated in Figure 3.



(a) Distribution of number of masks

(b) Distribution of areas of masks

Figure 4: Data statistics of ParaDLC-Bench

D.4 Data Statistics

To intuitively illustrate the data composition and task difficulty of ParaDLC-Bench, we conducted a multi-dimensional statistical analysis on the finalized dataset:

Distribution of Number of Masks: To highlight the core setting of "multi-target interaction," Figure 4 details the distribution of the number of tested masks contained within a single image. All images in this benchmark contain 2 or more masked regions, with the vast majority ranging from 2 to 4. Some extremely challenging samples even contain up to 8 masks. This high-density distribution severely tests the model's local perception and anti-interference capabilities within complex contexts.

Distribution of Areas of Masks: Figure 4 presents the area proportion distribution of the masked regions relative to the entire image. To fully test the model's local description capabilities at a fine-grained level, we intentionally retained a large number of micro-detail targets (small masks) that occupy an extremely small proportion of the image. The average mask area ratio is 0.07. This distribution highlights the importance of addressing small-scale and fine-grained understanding.

E Related Work

Diffusion Language Models. Diffusion Language Models (DLMs) have recently emerged as a promising alternative to autoregressive language modeling. Among them, masked diffusion models (You et al., 2025; Ye et al., 2025; Yang et al., 2025b), have shown particularly strong empirical performance. Building on this line of work, LLaDA scales masked diffusion language modeling to 8B parameters and demonstrates that diffusion-based language models can approach the performance of strong autoregressive LLMs, such as LLaMA3-8B, across a wide range of downstream tasks. Recent LLaDA series further scale the model size to over 100B parameters, with a more complex architecture (Mixture of Experts, MoE) and stronger reasoning capabilities. Meanwhile, several works also extend DLMs to multimodal settings, including visual understanding and generation (Yang et al., 2025b; Xin et al., 2025). LLaDA-V (You et al., 2025) adapts the visual instruction tuning paradigm to masked diffusion language modeling, while MMaDA (Yang et al., 2025b) extends DLMs to both generation and understanding. After that, several works explore AR-based adaptation by leveraging pre-trained AR knowledge (Cheng et al., 2025; Bie et al., 2025). There are also some works that accelerates DLM inference by reducing the number of denoising steps, through pseudo-trajectory distillation (Qian et al., 2026) or entropy-based multi-token decoding. Our

work mainly focuses on multimodal understanding tasks, where we build a stronger baseline that sets a new state of the art among diffusion VLMs.

Image Captioning in MLLMs. Most image caption datasets are used for multimodal pre-training tasks, which is the first stage of MLLMs for text and vision alignment (Li et al., 2025a). With the rapid development of MLLMs, this has become one of their key capabilities. Recent works explore various caption settings, including dense, detailed, grounded, and region-sensitive formats (Li et al., 2025b; Wang et al., 2025a). All previous solutions focus on AR-based image-to-text generation. Several multimodal DLMs can generate text in parallel. In contrast, we aim to explore multiple-object caption generation in parallel using DLMs, enabling parallel generation at both the sequence (object) and token levels. To our knowledge, we are the first to achieve this function.

Region Understanding of MLLMs. In contrast to conventional MLLMs that primarily emphasize image-level understanding without dedicated extraction of localized features, region-level understanding requires MLLMs to accurately model region-specific attributes, fine-grained visual details, and spatial relationships within designated areas. Consequently, enhancing regional perception has become a central research focus for advancing MLLMs' visual perception capabilities. Existing works have adopted three mainstream strategies to represent image regions-of-interest (RoIs), namely visual markers (Yang et al., 2023), bounding boxes (Zhang et al., 2023; Chen et al., 2023; Rasheed et al., 2024), and segmentation masks (Zhang et al., 2024b; Yuan et al., 2025a; Zhang et al., 2025a; Yuan et al., 2025b; Wang et al., 2025a; Zhou et al., 2026; Lian et al., 2025). These models are predominantly designed for single-region, single-caption generation, and nearly all adopt autoregressive (AR) LLMs as their core inference backbone, which restricts concurrent multi-region processing. In contrast, our approach leverages DLMs that offer inherent efficiency advantages, thereby enabling natural support for high-efficiency *parallel* region caption generation.

F Visualization and Qualitative Results

Parallel Perception Capabilities. In Figure 5, we present qualitative examples demonstrating that PerceptionDLM can simultaneously generate highly detailed and accurate descriptions for multiple masked regions within a single image through a single diffusion process. Whether dealing with tightly packed objects (e.g., the various ingredients in the bowl) or spatially distributed subjects (e.g., the man and his fishing rod), the model successfully captures fine-grained attributes such as color, texture, shape.

Comparison with Baseline Models. In Figure 6, we compare PerceptionDLM against existing strong baselines, including LLaDA-V, GAR-8B, and Gemini-3.1-Pro. When faced with adjacent and visually complex masks, PerceptionDLM effectively isolates region-specific visual context, producing descriptions that strictly align with the intended target regions. In contrast, baseline models frequently suffer from severe hallucinations and feature confusion. For instance, standard diffusion VLMs and AR-based region models may mistakenly assign the attributes of a neighboring object (e.g., colors, text, or parts) to the current target mask, or hallucinate entirely non-existent background elements. By comparison, PerceptionDLM significantly reduces such errors, effectively mitigating the severe cross-region interference observed in other models.

G Limitations and Future Works

While PerceptionDLM significantly improves multi-region perception efficiency, certain limitations remain. First, the actual inference speed is still bounded by the multi-step denoising process inherent to diffusion models. We take a first step toward mitigating this in Appendix C via pseudo-trajectory distillation (Qian et al., 2026), but the resulting acceleration still comes at a measurable cost in caption quality; distillation objectives tailored to region-conditioned generation remain an open direction. Second, eliminating attribute entanglement in extremely dense or semantically similar regions remains challenging; introducing explicit region contrastive losses or more fine-grained attention masking could further mitigate hallucinations. Finally, to bridge the performance gap in complex reasoning tasks compared to advanced autoregressive models, applying Reinforcement Learning (RL) techniques to unlock deep reasoning capabilities represents a critical direction for future exploration.

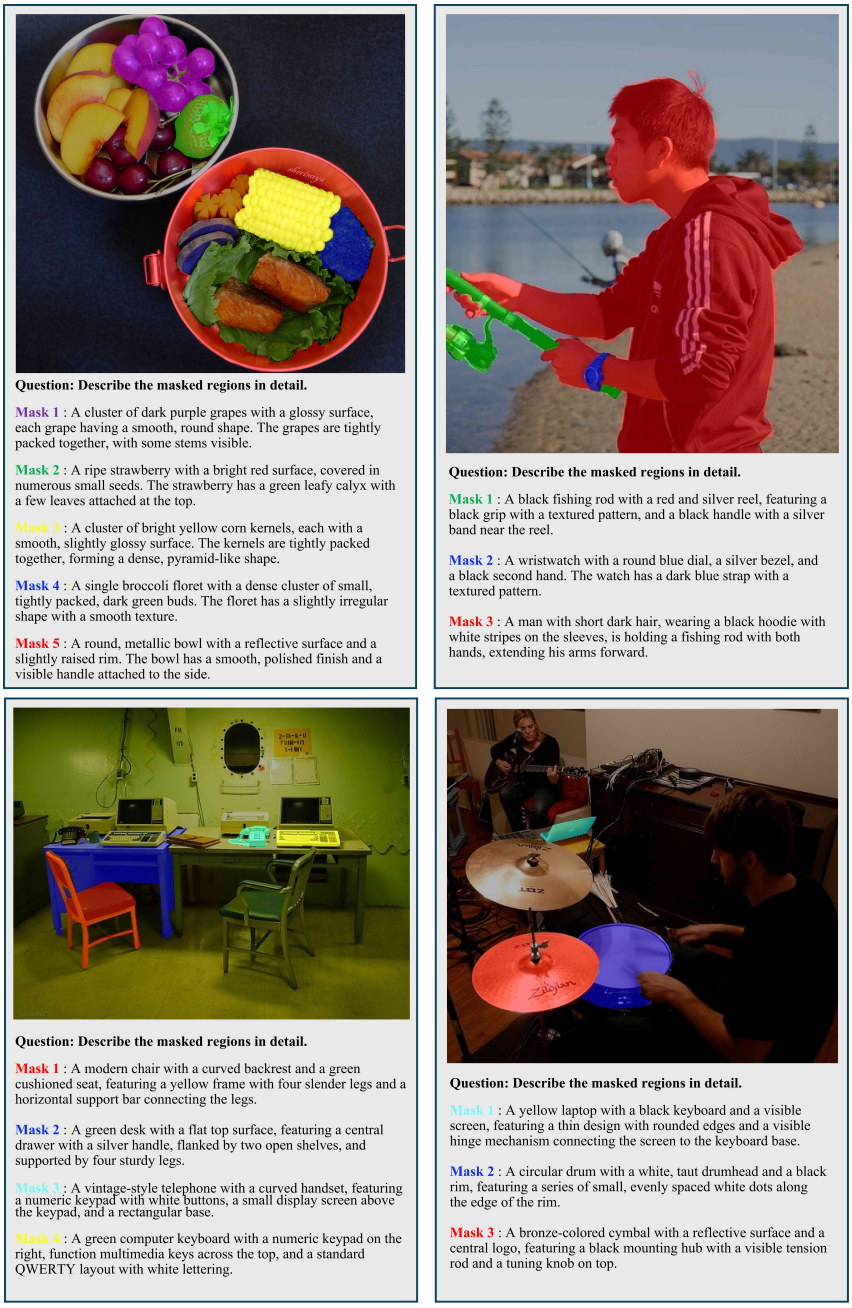


Figure 5: **Qualitative examples** of PerceptionDLM generating detailed descriptions for multiple masked regions in parallel.



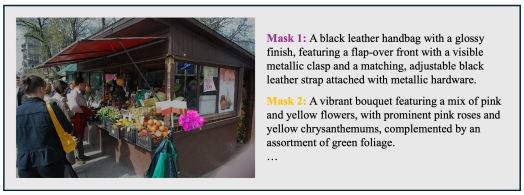
Model	Mask 1	Mask 2	Mask 3
PerceptionDLM	A person with dark hair tied back, wearing a red and black jacket, blue jeans, and red shoes, is captured mid-stride with their left leg forward and right arm slightly bent. The individual appears to be in motion, possibly walking or running, with their head turned slightly to the left, showcasing their profile.	A blue bus with a rear window displaying a white sign with red and black text, a metal ladder mounted on the back, and a set of red taillights on either side. The bus has a visible license plate area below the window, and the rear bumper is black with a tow hitch at the bottom center.	A dark-colored sedan with a visible rear end featuring a trunk with a centered license plate, a partially visible taillight on the left side, and a rear windshield with a defroster grid. The car has a prominent rear bumper with integrated reflectors and a visible rear wheel with a multi-spoke alloy design.
LLaDA-V	The target object in the segmented image is a person walking a dog. The person is wearing a red jacket and black pants, and the dog is attached to a leash. They are walking on a street, and the surrounding scene appears to in an urban environment with buildings, cars, and and pedestrians in the background.	The blue bus is a ladder-equipped public transport vehicle, with a ladder mounted on the top of the bus. The bus has a rounded front with a head bumper, a windshield at the front, and multiple windows on the side. The number \"713\" is visible on the the rear of the bus.	The target object in the image is a black sedan sedan. It is a behind the blue bus and a white bus, and it is specifically highlighted by the red bounding box in the image. The sedan has a distinct rectangular shape, a visible rear window, and is positioned on the right side of the image.
GAR-8B	A person with dark hair tied back, wearing a red jacket with a black collar, blue jeans, and red shoes, walking with a slight forward lean.	A blue bus with a rear emergency exit door, featuring a white and red stripe design on the lower half, a set of red taillights, and a visible license plate. The bus has a rear window with a wiper and a destination sign above it displaying text in a non-Latin script.	A dark-colored sedan with a glossy finish, featuring a rear spoiler and a visible rear windshield wiper. The vehicle has a trunk with a centered license plate, flanked by two red taillights that wrap around the sides, and a partially visible rear bumper with a reflective strip.
Gemini-3.1-pro	A person is captured mid-stride, viewed from a side profile. They are wearing a vibrant red, long-sleeved jacket with a high collar, paired with dark blue jeans and white sneakers. The individual has dark hair and is carrying a dark-colored backpack on their right shoulder.	This is the rear view of a tall bus with a vibrant, glossy blue-painted body. The vehicle has a rectangular shape with slightly rounded top corners. A thin, metal ladder is mounted vertically in the center of the rear panel, extending up towards the roof. Below the dark rear window, there is white text written in a non-Latin script. A white license plate with dark characters is visible on the lower left side, and a simple, thin white bumper runs along the bottom.	A glossy, dark blue sedan seen from a rear three-quarter angle. The car's paint has a shiny finish that reflects the bright ambient light. A horizontal chrome trim piece is visible on the trunk, positioned just above the red and white tail light assembly. Below the body, the black rubber of the rear tire is partially visible.

Figure 6: **Qualitative comparison of PerceptionDLM against existing baselines.** Correct attribute descriptions are highlighted in green, whereas hallucinations and errors are marked in red. PerceptionDLM effectively mitigates the severe cross-region interference observed in other models.

Failure Case Analysis. To better characterize the current limitations, we conduct a qualitative analysis of typical failure cases and group them into four representative categories, as illustrated in Figure 7:

- **(a) Cross-region attribute entanglement.** When multiple masks are spatially adjacent or semantically similar, the model occasionally leaks attributes (e.g., color, texture, or material) from a neighboring target into the current region's description. For example, in Figure 7(a), the caption of one mask absorbs the appearance of an adjacent object on the same market stall.
- **(b) Tiny or heavily occluded regions.** For extremely small or heavily occluded masks that occupy a negligible fraction of the image, limited visual evidence can lead to vague or partially incorrect descriptions. For example, in Figure 7(b), a small, low-resolution region is given an overly specific yet inaccurate description.
- **(c) Hallucination of typical-but-absent attributes.** Following the two negative types in ParaDLC-Bench, the model may fabricate parts commonly associated with the object category yet absent from the mask, or borrow attributes from co-occurring objects. For example, in Figure 7(c), it invents white laces, eyelets, and tread for a tiny, blurry shoe region that are indiscernible from the pixels.
- **(d) Fine-grained text.** The model can still struggle with reading small embedded text (OCR). For example, in Figure 7(d), it misreads the license plate as "201-VH" instead of the actual "267-JAN".

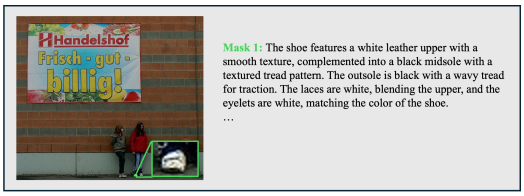
These observations are consistent with our quantitative findings and motivate the future directions discussed above.



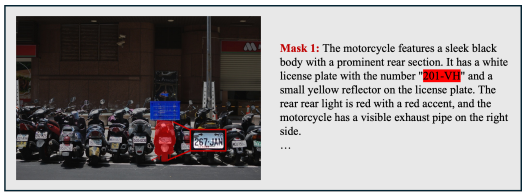
(a) Cross-region attribute entanglement



(b) Tiny or heavily occluded regions



(c) Hallucination of typical-but-absent attributes



(d) Fine-grained text

Figure 7: Representative failure cases of PerceptionDLM.

System Prompt for Question Generation

```
You are an expert dataset curator building an evaluation benchmark for detailed localized image captioning models. I
will provide you with a CROPPED image (showing the surrounding context) and a SEGMENTED image (showing only the
{category} of interest, with the background blacked out).
Your task is to analyze the {category} strictly based on the SEGMENTED image, using the CROPPED image only for
context.
```

Figure 8: System Prompt for Question Generation

Prompt for Positive Property Extraction

```
Based on the provided CROPPED and SEGMENTED images, perform the following tasks for the {category}:
1. List all distinct, visible parts of the {category}.
2. For the whole object and for each listed part, generate a comprehensive list of visual properties covering color,
shape, texture, materials, and size.
You MUST output your results strictly as a JSON array of tuples.
Each tuple has exactly 4 elements: [object_name, part_name, property_name, property_value].
Do not include any markdown formatting, explanations, or other text.
Example Output Format:
[
    ["cogi", "fur", "color", "brown"],
    ["sushi", "toping", "type", "sesame seeds"],
]
```

Figure 9: Prompt for Positive Property Extraction

Prompt for Negative Attribute Extraction

```
Based on the provided CROPPED and SEGMENTED images, perform the following tasks to identify potential AI errors
regarding the {category}:
1. Mislocalization Targets (Salient Negatives): List distinct objects or elements that are clearly visible in the
CROPPED image (surrounding context) but are completely OUTSIDE the masked area in the SEGMENTED image.
2. Hallucination Targets (Negative Parts): List parts, features, or properties that are commonly associated with a
typical {category}, but are explicitly NOT present, NOT visible, or occluded in this specific SEGMENTED image.
You MUST output your results strictly in the following JSON format.
Do not include any markdown formatting, explanations, or other text.
Example Output Format:
{{
    "mislocalization_targets": [
        "tape",
        "paper clips",
        "paintbrushes"
    ],
    "hallucination_targets": [
        "staple remover",
        "staple storage compartment"
    ]
}}
```

Figure 10: Prompt for Negative Attribute Extraction

System Prompt for Positive Question Formatting

```
You are an expert dataset curator. Your task is to convert verified property tuples into multiple-choice questions
for an image captioning evaluation benchmark. Follow the instructions and scoring rules exactly.
```

Figure 11: System Prompt for Positive Question Formatting

Prompt for Positive Question Formatting

```
Convert the following verified property tuples into a JSON array of multiple-choice questions for an image captioning
benchmark.
Target Category: {category}
Verified Properties: {tuples_str}
Instructions & Scoring Rules: Create a question for each property to test if a model correctly described that detail.
- Correct detail correctly mentioned: 1
- Factual error (wrong color/shape/etc.): -1
- Detail not mentioned, but the base object/part is mentioned: 0.5
- The base object itself is not mentioned: 0
Few-Shot Examples of Expected Output Format:
[
    {{
        "question": "Which of the following is applicable to the description?",
        "choices": [
            ["The material of the base plate is mentioned in the description and is metallic.", 1],
            ["The base plate or the stapler is not mentioned.", 0],
            ["The material of the base plate is mentioned in the description but is not metallic.", -1],
            ["The material of the base plate is not mentioned, but the base plate of the stapler is mentioned.", 0.5]
        ],
        "type": "positive"
    }},
    {{
        "question": "Which of the following is applicable to the description?",
        "choices": [
            ["The shape of the stapler is mentioned in the description and is rectangular.", 1],
            ["The stapler is not mentioned.", 0],
            ["The shape of the stapler is mentioned in the description but is not rectangular.", -1],
            ["The shape of the stapler is not mentioned.", 0]
        ],
        "type": "positive"
    }}
]
Output only the valid JSON array.
```

Figure 12: Prompt for Positive Question Formatting

System Prompt for Negative Question Formatting

```
You are an expert dataset curator. Your task is to convert verified negative targets into multiple-choice questions
for an image captioning evaluation benchmark. Follow the instructions and scoring rules exactly.
```

Figure 13: System Prompt for Negative Question Formatting

Prompt for Negative Question Formatting

```
Convert the following verified negative targets into a JSON array of multiple-choice questions for an image
captioning benchmark.
Target Category: {category}
Mislocalization Targets (Salient Negatives): {misloc_str}
Hallucination Targets (Negative Parts): {halluc_str}
Instructions & Scoring Rules:
- Salient Negatives use 2 choices: Mentioning the background object (-1) vs. Not mentioning it (1).
- Negative Parts use 3 choices: Mentioning the missing part (-1), Not mentioning the main object at all (0), vs.
Mentioning the main object but correctly omitting the missing part (1).
Few-Shot Examples of Expected Output Format:
[
    {{
        "question": "Which of the following is applicable to the description?",
        "choices": [
            ["The tape is mentioned in the description.", -1],
            ["The tape is not mentioned in the description.", 1]
        ],
        "type": "negative",
        "subtype": "salient negative"
    }},
    {{
        "question": "Which of the following is applicable to the description?",
        "choices": [
            ["The staple remover of the stapler is mentioned in the description.", -1],
            ["The stapler is not mentioned in the description.", 0],
            ["The staple remover of the stapler is not mentioned in the description.", 1]
        ],
        "type": "negative",
        "subtype": "negative part"
    }}
]
Output only the valid JSON array covering all provided targets.
```

Figure 14: Prompt for Negative Question Formatting