Falcon Perception: Natively Multimodal Autoregressive Perception Architecture

Phúc H. Lê Khắc[1*] Yasser Dahou[1*] Sanath Narayan[1*]
Ankit Singh[1] Wamiq Reyaz Para[1] Ngoc Dung Huynh[1]
Sofian Chaybouti[1,2] Hakim Hacid[1]

[1] Technology Innovation Institute, Abu Dhabi, UAE
[2] Tuebingen AI Center / University of Tuebingen, Germany

Abstract. Perception-centric systems are typically implemented with a modular encoder-decoder pipeline: a vision backbone for feature extraction and a separate decoder (or late-fusion module) for task prediction. We ask whether this separation is necessary for tasks such as open-vocabulary segmentation and OCR. We present Falcon Perception, a single, early-fusion dense Transformer that processes image patches, text, and task tokens in a single shared parameter space. Falcon Perception uses a hybrid attention mask: bidirectional for image tokens and causal for text/task tokens, and keeps dense outputs practical by emitting only a small number of task tokens per instance and decoding continuous spatial outputs with lightweight heads, enabling parallel high-resolution mask prediction. On SA-Co, Falcon Perception achieves 68.0 F_1, compared to 62.3 for a strong state-of-the-art promptable segmentation baseline. To measure compositional grounding beyond saturated referring benchmarks, we introduce PBench and show large gains on spatial and dense long-context regimes (e.g., 53.5 vs. 31.6 on spatial and 72.6 vs. 58.4 on Dense). Finally, we extend the same architecture to document OCR: a compact Falcon-OCR model with 300M parameters achieves 80.3% on olmOCR and 88.64% on OmniDocBench. Code: GitHub, PBench: HF Dataset, Blog: Project Blog.

1 Introduction

Perception systems for open-vocabulary segmentation and OCR are still largely built around a modular encoder–decoder recipe: a vision backbone produces features, and a separate decoder or late-fusion module turns them into task outputs [5,8,20,22,46]. This design has been effective, but it also encourages adding task-specific mechanisms (modality fusion, query matching, post-processing) that complicate scaling and limit effective feature fusion and learning between vision and text modality [4,27,33,38].

This raises a fundamental question: *do dense grounding systems actually need an encoder–decoder split to "see" and to "predict"?* If not, what should replace the standard late-fusion interface so that language and visual features can interact from the first layer, especially for compositional prompts involving spatial constraints, and relations. What is the right output interface for dense perception, where the number of instances can range from zero to hundreds, without making decoding prohibitively expensive [8,9,23]? And finally, how should we benchmark progress once standard

* Equal contribution.
Corresponding author: yasser.dahou@tii.ae

referring datasets saturate, in a way that isolates these compositional capabilities and stresses long-context crowded scenes?

In this paper, we study these questions and introduce Falcon Perception, a single early-fusion dense Transformer unifying the image patches, text, and task tokens with a hybrid attention pattern (bidirectional for image tokens, causal for prediction tokens), enabling a single set of weights to perform both perception and prediction (Figure 2). We also introduce PBench, a diagnostic benchmark targeting compositional grounding (attributes, spatial, relations) and dense long-context regimes, and extend the same backbone to document OCR with a compact Falcon Perception-OCR model (300M), evaluated on standard OCR benchmarks [42, 44]. Specifically, our contributions are:

- We introduce **Falcon Perception**, a 600M single early-fusion Transformer reaching competitive results on open-vocab segmentation.
- We propose an efficient dense output interface that retains autoregressive flexibility while decoding continuous spatial outputs with lightweight heads and parallel high-resolution mask prediction.
- We introduce **PBench**, a benchmark for compositional prompts and dense long-context evaluation, enabling capability-level analysis beyond saturated referring benchmarks.
- We demonstrate generality by extending Falcon Perception to OCR: **Falcon-OCR** (300M) achieves strong results on document OCR benchmarks.

2 Related Work

Traditional closed-vocabulary detection models [5, 16] utilize a single encoder to predict a fixed set of categories, while extending these systems to the open-vocabulary regime requires bridging visual and semantic domains. Methods such as SAM 3 [4] and OWL-ViT [37] address this by integrating a text encoder that generates embeddings for arbitrary queries, often relying on CLIP-based [45] pretraining to align image and text in a shared embedding space for zero-shot detection. Pix2Seq [8] formulates detection as a sequence-to-sequence task by serializing outputs into discrete tokens trained with a next-token prediction objective. Pix2SeqV2 [9], Florence [55], and UnifiedIO [9] extend this paradigm to multiple vision tasks through adaptive prompting. Moreover, recent generalist VLMs (*e.g.*, Qwen3-VL [57]) that demonstrate broad capabilities and Moondream [24, 40] that focus on small, deployment friendly VLMs are capable of box prediction.

Furthermore, Optical Character Recognition (OCR) is naturally a sequence-to-sequence perception task requiring long-form text generation. Recent approaches [28, 44] adopt VLM architectures to improve efficiency and accuracy. However, all these methods retain a modular encoder-decoder architecture, in which a vision encoder processes the image and a language decoder generates predictions. Although this modular design achieves strong results on tasks like image captioning and VQA [2, 10], it is prone to hallucinations [52] and struggles with fine-grained perception tasks [18]. We challenge the assumption that different modalities require substantially different modules and complex training strategies to bridge the "modality gap" [30], even for perception tasks such as open-vocab segmentation, and OCR that demand high-fidelity visual conditioning while generating text tokens.

Different from the modular VLMs, LIMoE [41] and Fuyu [1] explore early-fusion (encoder-free) architectures by projecting image patches with a linear layer and feeding them, together with text embeddings, into a shared Transformer. This is in line with [17, 36], which suggest that networks trained on different modalities may converge to a shared representation space. Subsequent works

Table 1: Taxonomy of our PBench referring expression segmentation benchmark. Each sample is assigned a single level label by construction, enabling per-level capability profiling rather than a saturated aggregate score.

	Level 0	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Rules	Whole objects; clear boundaries; bbox-drawable; no abstract concepts.	**Attribute:** color, size. **State:** old/new, broken. **Subtype:** sedan/SUV. **Function:** open/closed.	**Brand:** Nike, Coca-Cola. **Variant:** Diet vs Reg. **Store:** Starbucks. **Sign:** sale tag.	**Rel:** left/right, above. **Depth:** front/back. **Order:** 3rd from left. **Cont:** inside/outside.	**Act:** holding, wearing. **Func:** key for door. **Comp:** tallest/smallest. **Phys:** resting on.

Ex.	`car`
	`person`
	`tree`
		`red car`
		`broken fence`
		`open door`
			`Diet Coke`
			`Nike shoes`
			`Exit sign`
				`car on left`
				`3rd window`
				`person in back`
					`holding bottle`
					`pulling trailer`
					`tallest tree`

reintroduce modality-specific components: MoMa [7,31] employs MoE layers [26], while [29,35] add modality-specific projections and normalization, retaining shared global attention. Different from the aforementioned works, we pursue a strictly shared early-fusion design, where a single linear-layer tokenizes image patches and the Transformer comprising the majority of parameters is fully shared across text and vision. Unlike general VLMs built on pretrained LLMs, we adopt a vision-first approach tailored to perception tasks, initializing from a strong image encoder where language comes late.

The closest to our early-fusion, natively multimodal model for perception task is PixelSAIL [59], built on top of the encoder-free VLM SAIL [25]. In contrast to PixelSAIL that distills a high-resolution image feature upsampler, we use a model-agnostic learned feature upsampler [54]. This enables us to freeze the upsampler module during training, thus reducing memory requirements and allowing us to work directly on high-resolution segmentation masks. Conscious of our model's size, we also focus only on detection and segmentation as a 'Chain-of-Perception' sequence, not supporting more complex tasks such as VQA and visual prompting, resulting in a lean, simple yet still achieving state-of-the-art perception results.

3 PBench: Perception Benchmark

Standard benchmarks like RefCOCO, RefCOCO+ [58], and G-Ref [21] suffer from performance saturation and a critical lack of semantic granularity, conflating diverse challenges into a single scalar metric. To rigorously evaluate Falcon Perception and understand its capabilities, we introduce a hierarchical evaluation benchmark with samples organized into five distinct levels of complexity. Designed to disentangle the capabilities required for open-vocabulary segmentation, this allows us to track progress on capabilities effectively.

The benchmark contains approximately ∼5k samples across Levels 0–4, and an additional ∼400 samples dedicated to crowdedness stress testing. Each sample is assigned a *single* level label by construction, and prompts are intentionally written to target one dominant capability while avoiding cross-level cues (*e.g.*, OCR prompts avoid spatial qualifiers such as "left/right", and spatial prompts avoid in-image text disambiguators). This yields a capability profile rather than a single scalar score. Please refer to Table 1 for the levels definition.

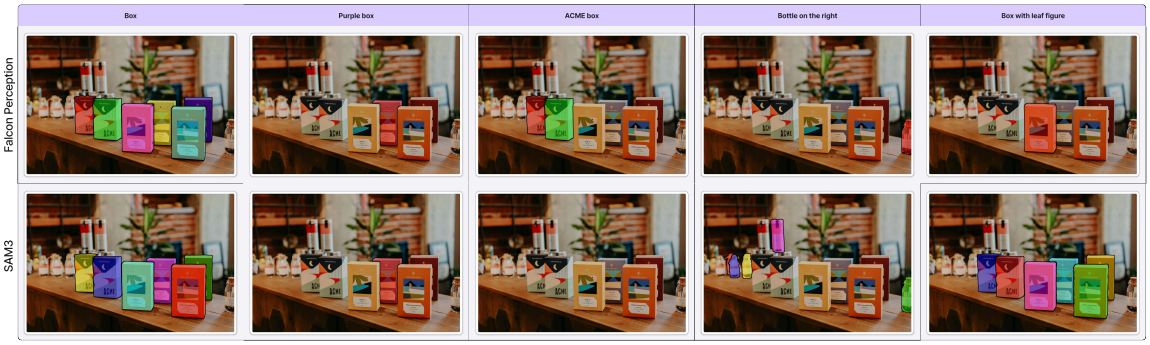


Fig. 1: Prompt progression across complexity levels. We vary the text prompt to isolate the capabilities targeted by Levels 0–4. Each panel shows the predicted masks for the same scene under progressively more specific prompts: from generic object classes (Level 0, *e.g.* "box"), to attribute binding (Level 1, *e.g.* "purple box"), to OCR-based identification (Level 2, *e.g.* "ACME box"), and then spatial/layout constraints (Level 3, *e.g.* "bottle on the right") and fine-grained relational descriptions (Level 4). Falcon Perception remains stable as prompts become more compositional and text-dependent, while SAM 3 starts to fail on OCR-driven queries and higher-level constraints, either returning no masks or segmenting visually plausible but semantically wrong instances.

How to interpret the levels: The levels are not a subjective notion of "difficulty"; rather, each level isolates a dominant capability. Levels 0-1 primarily probe open-vocabulary recognition and attribute binding. Level 2 isolates OCR-aware disambiguation, where in-image text serves as the deciding signal between near-identical instances. Level 3 isolates spatial grounding and layout understanding, testing whether the model forms a coherent 2D scene map. Level 4 isolates relational grounding where the target is defined through interaction rather than appearance. We report per-level performance in addition to an overall average, yielding a capability profile that reveals *where* a model fails (*e.g.*, collapsing at OCR or relations).
Dense split: We also evaluate robustness to scene density by measuring performance as a function of instance count K (up to $K \approx 600$). Testing whether the autoregressive Chain-of-Perception interface can maintain precise localization and avoid drift or duplication in highly crowded environments.

4 Method

The simplicity of Falcon Perception lies in unifying visual perception and language understanding within a single transformer backbone f_θ, deviating from the standard encoder-decoder pipeline. We process raw pixel patches and text tokens with the same set of weights, enabling deep, early fusion and joint training across modalities. Clearly, however, applying pure language modeling for dense perception tasks (*e.g.*, generating hundreds of polygon coordinates per instance) is computationally expensive. To resolve this, we adopt a hybrid approach: the backbone autoregressively predicts special object tokens (*e.g.*, `<seg>`), whose hidden states then act as queries for a lightweight, specialized head. This design preserves the simple interface of a sequence model while enabling the efficient, parallel generation of dense binary masks.

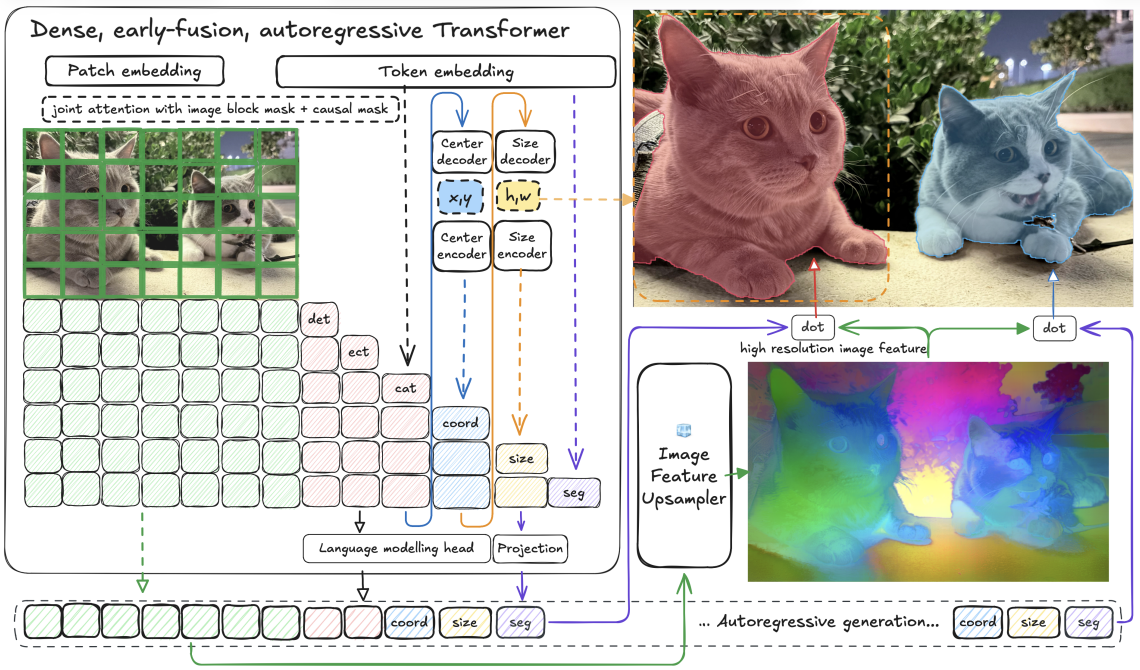


Fig. 2: Falcon Perception: A single autoregressive Transformer processes a unified sequence of image patches, text, and task tokens. The model predicts object properties in a fixed order: `<coord>` → `<size>` → `<seg>`. Bounding-box coordinate and size tokens are decoded via specialized heads and reinjected as Fourier features to condition subsequent steps. High-resolution segmentation masks are generated by a dot product between the `<seg>` token of each instance and the upsampled image features, leveraging the early-fusion backbone for instance-aware localization. Visual data flow is shown in green, coordinates in blue, size in orange, and segmentation in purple.

4.1 Autoregressive backbone

We formulate dense perception as a conditional generation task. Given an image $I \in \mathbb{R}^{H \times W \times 3}$ and a text prompt $\mathcal{P} = \{t_1, \ldots, t_L\}$, the model must predict a set of K objects $\{O_k\}_{k=1}^{K}$. Each object is defined by a tuple $O_k = (c_k, s_k, m_k)$ representing its center coordinates, size, and binary segmentation mask. To solve this, we serialize the output into a structured sequence, imposing a "Chain-of-Perception" order: $c_k \rightarrow s_k \rightarrow m_k$. This coarse-to-fine curriculum stabilizes training by forcing the model to resolve spatial ambiguity before committing to pixel-level details. The resulting sequence format is `[Image] [Text] <coord> <size> <seg> ... <eos>`, as illustrated in Figure 2. Predicting coordinates and size *before* segmentation forces the model to resolve instance ambiguity (*e.g.*, "which cat?") and spatial extent first. This provides a strong conditioning signal for the subsequent segmentation token `<seg>`, allowing it to focus on fine-grained pixel refinement rather than instance localization, eliminating the need for a heavy mask decoder.

Input Representation: To process these heterogeneous inputs, we flatten the image into N patches and project them into visual embeddings $V \in \mathbb{R}^{N \times d}$. Similarly, the text prompt is mapped to L text embeddings $T \in \mathbb{R}^{L \times d}$. The unified sequence $X \in \mathbb{R}^{S \times d}$ is formed by concatenating these with the task tokens for the K objects, where $S = N + L + 3K$ is the total sequence length and d is the

hidden dimension:

$$X = [\underbrace{v_1, \ldots, v_N}_{\text{Visual Embeddings}}, \underbrace{t_1, \ldots, t_L}_{\text{Text Embeddings}}, \underbrace{e_{\texttt{<coord>}}, e_{\texttt{<size>}}, e_{\texttt{<seg>}}, \ldots}_{\text{Task Embeddings}}] \tag{1}$$

The model processes this sequence autoregressively. The backbone f_θ is a standard dense Transformer with N_ℓ layers. Processing this sequence requires a specialized attention strategy. Standard causal masking is suboptimal for visual data, as image patches are inherently 2D thus benefit from bidirectional context. We therefore employ a hybrid masking strategy: image tokens attend to all other image tokens (bidirectional), while text and task tokens attend to all image tokens but only to preceding text/task tokens (causal). This allows a single set of weights to act simultaneously as a bidirectional vision encoder and an autoregressive language decoder. Unlike Prefix-LM masking [3], we do not allow the prefix text tokens to attend to each other bidirectionally.

4.2 Specialized Heads

While keeping a single Transformer backbone to maximize interaction between tokens of all modalities, we employ specialized encoders and decoders to inject extra inductive biases for the spatial task at the input and output layers.

Coordinate & Size Encoders using Fourier Features: Standard tokenization of coordinates (*e.g.*, 1000 bins) suffers from limited precision and *spectral bias*, the tendency of neural networks to prioritize low-frequency functions, resulting in imprecise spatial grounding. To address this, we adopt a Fourier Feature mapping strategy [50] that maps input coordinates or sizes $\mathbf{c}, \mathbf{s} \in [0, 1]^2$ to a higher-dimensional feature space using a random Gaussian matrix $\mathbf{B} \in \mathbb{R}^{d/2\times 2}$: $\gamma(\mathbf{c}) = [\cos(2\pi\mathbf{B}\mathbf{c}), \sin(2\pi\mathbf{B}\mathbf{c})]$. Each entry in the matrix $\mathbf{B}$ is drawn from a normal distribution $N(0, \sigma)$, where the scale σ (set to 10.0) controls the bandwidth of the kernel, tuning sensitivity to spatial detail. This projection allows the model to ingest precise object coordinates over a rich frequency spectrum. These continuous embeddings are then linearly projected to the transformer dimension and injected directly into the input stream, replacing the static `<coord>` or `<size>` token embeddings. For decoding, we use a simple MLP to project the hidden state into $2 \times N_{bins}$ logits.

Segmentation (Upsampling + Dot Product): Generating a high-resolution mask $m \in \mathbb{R}^{H\times W}$ from a single token is challenging. Standard approaches, such as Mask2Former [11], require complex Hungarian matching to resolve instance ambiguity. We simplify this by leveraging our unified architecture. Since the `<seg>` token is generated after the coordinates and size, the model has already resolved the object identity. Furthermore, due to our early fusion, its hidden state h_{seg} has direct access to the global visual context. Thus, the mask can be generated via a simple dot product, without separate mask queries.

To recover spatial detail, we employ an image-aware feature upsampler $\mathcal{U} : (V_{out}, I) \to V^*$ [54]. This model-agnostic learned module formulates upsampling as a cross-attention operation where the query is derived from the high-resolution input image $I \in \mathbb{R}^{H\times W\times 3}$ (to guide boundaries), and the key/value are derived from the backbone's visual features $V_{out} \in \mathbb{R}^{N\times d}$. This allows the model to project semantic features onto the high-resolution pixel grid, producing an enhanced feature map $V^* \in \mathbb{R}^{H\times W\times d}$ without any further feature transformation. The final mask $\hat{m}$ is generated by computing the dot product between these features and the projected hidden state of the segmentation token h_{seg}:

$$\hat{m} = \sigma(V^* \cdot \text{Proj}(h_{seg})^T) \tag{2}$$

Because h_{seg} is generated autoregressively and has access to the full early-fusion image context V_{out}, it can encode any necessary instance-specific information. Consequently, the segmentation task is reduced to a straightforward dot product between its hidden state and the high-resolution feature map, creating a sharp, instance-specific segmentation mask. This result in a lightweight and simple segmentation head, making the architecture simple to implement and scale.

5 Training details

We adopt a two-stage approach to effectively learn dense perception capabilities. In the first stage of Multi-Teacher Distillation, following [6], we distill knowledge from strong vision teachers (DINOv3 [49] and SigLIP2 [51]) into our unified dense backbone. This provides a robust initialization for visual features. In the next stage, we train the model on approximately 700 GT of multimodal data using our specialized "Chain-of-Perception" format (`<coord><size><seg>`). This stage fine-tunes the backbone and heads to perform autoregressive dense prediction tasks. We train on a curated dataset of 54M images (195M expressions) constructed via a multi-stage pipeline: **Clustering** → **VLM Listing** → **Ensemble Consensus** → **Human Verification**. This ensures broad semantic coverage across complexity levels (0–4), with hard disagreements routed to human annotators to prevent bias. We further mix in public vision datasets and text-only corpora [43] to maintain general capabilities (see supplementary Section B).

5.1 Architecture details

Beyond the core backbone and specialized heads, several key architectural decisions are needed for handling dense spatial information effectively. We detail these choices below.

3D Rotary Positional Embeddings (Sequence + Space): To preserve the spatial structure of the input within the flattened sequence X, we inject positional information directly into the attention mechanism. Standard 1D RoPE treats the image as a linear sequence, destroying the 2D grid relationships critical for dense prediction. We instead decompose the head dimension d_{head} into two distinct subspaces: a sequential component and a spatial component. The first half of the head dimension ($d_{head}/2$) encodes the 1D sequence index t using standard RoPE, capturing the causal order of the text tokens and the relative position of patches in the flattened sequence. The second half ($d_{head}/2$) encodes the 2D grid position $\mathbf{p} = (x, y)$ using **Golden Gate RoPE (GGRoPE)** [56]. Unlike standard Axial RoPE, which bias attention to orthogonal x and y axes, GGRoPE rotates each dimension pair j based on the projection of the spatial position $\mathbf{p}$ onto a unique direction vector $\mathbf{u}_j$: $\theta_j = \omega_j(\mathbf{p} \cdot \mathbf{u}_j)$ where the direction vectors $\{\mathbf{u}_j\}$ are sampled uniformly from the unit circle using the Golden Ratio ϕ. This mechanism allows the attention heads to attend to relative positions along arbitrary 2D angles, producing isotropic attention maps that are robust to rotation and aspect ratio variations. Also, this spatial rotation is applied sparsely only to valid visual tokens, ensuring computational efficiency and preventing interference with text embeddings does not have spatial coordinates.

Native Resolution Handling: To avoid distorting aspect ratios and degrading performance on small objects. We process images at their native resolution and aspect ratio (up to a maximum token budget). This introduces high variance in sequence lengths: a 256×256 image yields ~ 256 patches, while a 768×768 image yields $\sim 2,304$. To handle the resulting variance in token counts efficiently, we employ a scatter-and-pack strategy: we patchify the image and discard all padding tokens *before* the transformer, keeping only the valid visual patches. These variable-length sequences are then

Table 2: AdamW *vs.* Muon

Optimizer	PBench-*Det*	SaCo-*Det*
AdamW	56.4	49.0
Muon	**57.7**	**53.8**

Table 3: Gram loss impact

Feature Reg.	PBench-*Seg*	SaCo-*Seg*
No Gram	52.7	51.1
With Gram	**53.8**	**52.6**

Table 4: Query masking

Ordering	PBench-*Det*	SaCo-*Det*
Full AR	53.2	53.1
Query Masking	**54.2**	**53.3**

packed into a single sequence of fixed length C_{max}. We utilize FlexAttention to implement the appropriate masking, ensuring that self-attention is restricted within each image sample's boundaries. This maximizes GPU utilization by processing only valid data, but results in a variable number of images per batch across ranks. We address the resulting optimization instability via a global loss balancing strategy [6].

5.2 Design choices

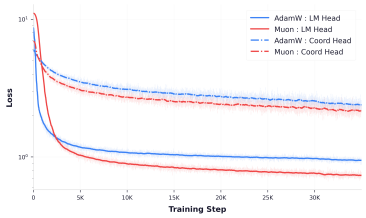
Training a native early-fusion model with multiple objectives introduces design challenges. Here, we detail the key sweeps conducted to finalize the design. Settings for each experiment were kept constant across variants. For efficiency, we perform ablations only on the detection task, except for feature regularization, where segmentation is used to better analyze retention of high-quality image features. A sample with multiple expressions is represented in a unified format:
`<image>expr1<present><coord><size><seg><eoq>expr2<absent>`$\cdots$`<eos>`,
where `<present>` and `<absent>` indicate whether such a referring expression object is present or not in the image.

***(i)* Loss Balancing across Ranks:** In our multimodal training with packed sequences, the number of valid tokens varies significantly across data-parallel ranks due to differences in image size, the number of expressions, and the number of instances per packed sample. We identified that standard FSDP gradient averaging biases the loss towards ranks with fewer tokens. We implemented a global normalization strategy in which the loss is first aggregated across ranks and then scaled by the *global* average number of tokens (N_{total}/R) rather than the local rank count. Please check supplementary section A.3 for details.

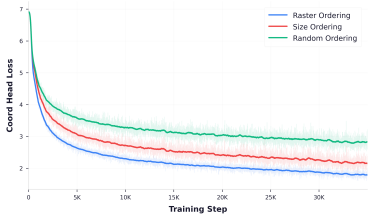
***(ii)* Optimizer Choice: Muon *vs.* AdamW:** We compared standard *AdamW* [34] against *Muon* [19], to validate the effectiveness of Muon in the multi-head and multi-loss setting. Fig. 3a shows the training loss curves for the Muon and AdamW optimizers on the language modeling and coordinate heads. We use the Muon's update rule from Kimi Moonlight [32] to share the optimal learning rate for both optimizers, and observe that employing the Muon optimizer leads to lower training losses across the different heads, yielding improved performance on both PBench and SaCo benchmarks in terms of macroF1 (Tab. 2).

***(iii)* Feature Regularization (Gram Loss):** We introduce the Gram loss, similar to DINOv3 [49], as a regularization to maintain the structural integrity of the visual feature space. This enforces that the correlation matrix (Gram matrix) of the student's patch features matches that of a frozen teacher model. From the ablation experiment in Tab. 3, we observe that training with the Gram loss indeed results in better performance on both benchmarks since it enables better image feature retention that is required for segmenting the instances, compared to learning the task without the Gram regularization.

***(iv)* Attention and Causality:** We investigated the attention mechanism between queries. *Full AR* vs. *No Inter-query Attn:* We compared full causal attention (where queries attend to the image features in addition to the previous queries and their corresponding instances) against a masked



(a) Optimizer comparison



Ordering	PBench-*Det*	SaCo-*Det*
Random	52.2	46.3
Size	57.7	53.8
Raster	59.3	56.2

(b) Ordering of instances

Fig. 3: (a) Muon optimizer results in lower training losses for the language and coord heads, in comparison to AdamW, leading to improved performance (see Table 2). (b) Raster ordering of instances results in lower coordinate head loss and better performance on both benchmarks, compared to random and size orderings.

approach, where queries attend to the image but not to each other, assessing the trade-off between relational reasoning and computational efficiency. Full AR enables in-context learning among the queries, where the model can learn to skip predicting the positions occupied by instances from previous queries. In contrast, inter-query masking attention forces the model to focus on image content to predict instances for a given query. Although the ablation in Tab. 4 shows a marginal advantage for the query masking variant, since both the strategies rely on complementary information for prediction, we employ them both in different stages of our model training, as detailed in Sec. 5.3.

***(v)* Mask Ordering:** When multiple objects are present, the order of target sequences matters. We compared *random*, *size* (largest-to-smallest instances), and *raster* ordering to determine whether a canonical curriculum improves object recall. We observe that raster ordering converges faster to a lower loss for the coord head among the three variants (Fig. 3b). We hypothesize that raster is better because it enables our model to search for objects in a clear order (which is not possible with the other two variants) and to learn faster with less ambiguity. And since the coordinate token is predicted first for an instance, lower coord loss (*i.e.*, a better learned coord head) directly translates to an improved detection performance in terms of macroF_1 on both PBench and SaCo benchmarks.

5.3 Training Recipe

Following multi-teacher distillation, we train Falcon Perception for ∼700 gigatokens (GT) with a three-stage curriculum that transitions from scene-level listing to the final independent-query setting. We keep a 1:1 positive/negative (present *vs.* absent expressions in the image) ratio throughout to reduce hallucinations. Unless stated otherwise, we train a 600M-parameter model with Muon. Images are processed up to 1024^2 resolution with a patch size of 16.

Stage 1: In-context listing (450 GT). Full autoregressive training over the entire text sequence (including expressions) with inverse LR decay 4e-4 → 1e-4 and max 100 masks/expression.

Stage 2: Task alignment (225 GT). We match inference by (i) *query-masked attention* (blocks cannot attend across query blocks) and (ii) *prompt loss masking* (no loss on expression tokens), employing inverse LR decay 1e-4 → 4e-6.

Stage 3: Long-context finetuning (10 GT). Short adaptation to crowded scenes with max 600 masks/expression and fixed LR 1e-6.

Fig. 4: Qualitative comparison across PBench complexity levels. Each column shows a different image and a prompt written to target one dominant capability (Levels 0–4; Table 1). The top row shows Falcon Perception predictions and the bottom row shows SAM 3. Falcon Perception produces cleaner, more instance-specific masks as prompts become more compositional (OCR, spatial, relational), while SAM 3 more often fragments the target or returns extra/incorrect masks on higher-level prompts.

6 Results

Metrics: We follow SAM3 [4] and report both Positive Micro F_1 (pmF_1) and Image-Level MCC (IL_MCC) as the primary metrics for open-vocabulary perception-centric segmentation. We also consider Macro F_1 per sample (F_1). Additional details on the metrics are given in the supplementary A.2.

6.1 Main Results

We evaluate Falcon Perception on: (1) The SA-Co benchmark for open-vocabulary segmentation, (2) Our new Perception Benchmark (PBench) for fine-grained capability analysis, and (3) The standard RefCOCO suite. To enable a fair comparison on segmentation metrics, we equip detection-only models (Qwen3-VL [57], Moondream [24, 40]) with SAM [22], using their output boxes as prompts to generate pixel masks.

SA-Co (open-vocabulary segmentation): Table 5 presents our performance on the SA-Co benchmark, showing that Falcon Perception is competitive on SA-Co and is particularly strong on *mask quality* as measured by Macro-F_1. On average, Falcon Perception reaches **68.0** $\mathbf{F_1}$ compared to **62.3** for SAM3, and it improves F_1 on most splits (*e.g.*, *Food&Drink:* **70.3** ***vs.*** **58.1**, *Sports:* **75.2** ***vs.*** **71.2**, *Attributes:* **79.3** ***vs.*** **71.1**). Falcon Perception lags behind SAM 3 in *presence calibration*, and this gap is the main driver of lower overall robustness on splits with many hard negatives (*e.g.*, *Average MCC:* **0.64** ***vs.*** **0.82**, *Wiki-Common MCC:* **0.38** ***vs.*** **0.66**).

PBench (capability profiling): While SA-Co studies performance across seven domains, PBench isolates *what capability is required* by construction (Levels 0–4 & Dense).

Table 5: SA-Co Benchmark Results. Comparison of Falcon Perception against state-of-the-art open-vocabulary segmentation models. We report Macro F_1, Positive Micro F_1 (pmF_1), and Image-Level MCC across all splits.

Model	Average			Metaclip			SA-1B			Crowded			Food&Drink			Sports Equip.			Attributes			Wiki-Common		
	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC	F_1	pmF_1	MCC
Baselines																								
gDino-T [33]	-	16.2	0.15	-	13.9	0.21	-	15.4	0.20	-	3.4	0.08	-	9.8	0.10	-	11.2	0.10	-	47.3	0.29	-	12.1	0.06
OWLv2* [38]	-	42.0	0.57	-	34.3	0.52	-	26.8	0.50	-	30.7	0.51	-	49.4	0.65	-	56.2	0.64	-	56.2	0.63	-	40.3	0.54
OWLv2 [38]	-	36.8	0.46	-	31.3	0.39	-	21.7	0.45	-	24.8	0.36	-	47.9	0.51	-	47.0	0.52	-	48.2	0.54	-	36.6	0.49
LLMDet-L [15]	-	27.3	0.21	-	19.4	0.23	-	22.8	0.23	-	13.7	0.18	-	29.1	0.19	-	25.3	0.17	-	57.1	0.39	-	23.3	0.16
APE-D [48]	-	36.9	0.40	-	30.1	0.42	-	10.0	0.22	-	20.3	0.35	-	45.0	0.51	-	56.5	0.56	-	57.3	0.47	-	39.5	0.41
DINO-X [47]	-	55.2	0.38	-	49.2	0.35	-	40.9	0.48	-	37.5	0.34	-	61.7	0.49	-	69.4	0.41	-	**74.0**	0.42	-	53.5	0.35
Vision-Language Models																								
Gemini 2.5 [12]	-	46.1	0.29	-	33.8	0.29	-	32.1	0.41	-	30.3	0.27	-	59.5	0.33	-	53.5	0.28	-	63.1	0.30	-	50.3	0.25
Qwen3-VL-2B [57]	50.3	36.8	0.65	45.4	31.9	0.63	46.4	31.5	0.71	26.4	17.1	0.62	53.4	36.7	0.70	60.5	51.1	0.75	68.4	54.5	0.61	51.7	35.1	0.52
Qwen3-VL-4B [57]	59.0	46.1	0.69	54.9	41.3	0.65	55.9	41.3	0.77	37.5	27.2	0.67	58.7	40.7	0.77	68.1	61.1	0.77	75.5	64.1	0.68	62.4	46.9	0.54
Qwen3-VL-8B [57]	56.4	41.1	0.79	51.4	36.2	0.76	53.3	34.8	0.82	32.2	19.9	0.77	56.3	37.4	**0.85**	67.4	54.9	0.87	74.4	62.3	**0.77**	59.7	42.2	0.77
Qwen3-VL-30B [57]	62.8	49.9	0.64	59.0	45.5	0.59	57.8	44.9	0.73	40.7	31.1	0.57	63.1	41.4	0.70	73.7	66.3	0.75	78.6	69.3	0.63	**66.8**	50.7	0.49
Moondream2 [24]	62.5	53.5	0.43	56.5	45.7	0.42	54.8	41.5	0.58	43.8	40.5	0.45	67.3	55.0	0.43	72.9	68.9	0.49	76.7	69.6	0.38	65.2	53.3	0.43
Moondream3 [40]	58.0	47.8	0.45	53.2	40.2	0.43	51.9	36.8	0.61	35.6	30.6	0.44	59.5	46.2	0.45	69.2	63.9	0.51	75.4	67.6	0.42	61.0	49.5	0.44
SAM 3 [4]	62.3	**66.1**	**0.82**	56.0	**58.6**	**0.81**	**68.3**	**62.6**	**0.86**	**62.7**	**67.7**	**0.90**	58.1	**67.3**	0.79	71.2	**73.8**	**0.89**	71.1	72.0	0.76	49.0	60.9	0.66
Ours: Falcon Perception	**68.0**	62.1	0.64	**63.8**	53.2	0.67	64.7	49.7	0.76	59.1	57.9	0.68	**70.3**	66.2	0.61	**75.2**	71.3	0.75	**79.3**	72.2	0.60	63.9	**64.3**	0.38

Table 6 shows that Falcon Perception stays robust as prompts become more compositional. Compared to SAM 3, gains are modest on Level 0 (simple classes), but become substantial on higher levels where semantic understanding is critical: **+9.2** on Level 1 (attributes), **+13.4** on Level 2 (OCR-guided), and a remarkable **+21.9** on Level 3 (spatial understanding). This confirms that while SAM 3 excels at promptable segmentation, it is not designed to resolve complex semantic ambiguities (Levels 2–4) without an external reasoning module. Crucially, Falcon Perception (0.6B) is highly competitive with far larger VLMs. It outperforms Moondream2-2B and Moondream3-9B on almost all complex reasoning tiers (L1–L4) and the Dense split, despite being 3×–15× smaller. Against the Qwen3-VL family, Falcon Perception matches or exceeds the 8B model on spatial (L3) and relational (L4) tasks, and even surpasses the massive Qwen3-VL-30B on the Dense split (**72.6** *vs.* 8.9). This highlights the efficiency of our native dense architecture: generalist VLMs struggle to maintain precise localization in crowded scenes (Dense split), often hallucinating or merging instances, whereas Falcon Perception's Chain-of-Perception interface remains stable as the number of targets increases.

Table 6: PBench & RefCOCO Results. Top: PBench performance (macro F_1) across complexity levels and crowdedness stress test. **Bottom:** RefCOCOm performance (macro F_1) on validation set.

Benchmark	Qwen3-VL-2B	Qwen3-VL-4B	Qwen3-VL-8B	Qwen3-VL-30B	Moondream2-2B	Moondream3-9B	SAM 3-0.9B	Falcon Perception-0.6B
PBench								
L0: Simple objects	54.1	67.3	65.6	**69.2**	68.0	63.2	64.3	65.1
L1: Attribute	50.1	63.9	65.6	**68.8**	58.4	59.6	54.4	63.6
L2: OCR guided	41.2	58.2	58.2	**61.2**	46.6	41.8	24.6	38.0
L3: Spatial understanding	35.1	49.0	49.1	52.9	43.8	45.4	31.6	**53.5**
L4: Relation binding	36.7	48.3	50.9	**55.2**	36.9	42.3	33.3	49.1
Dense	4.6	8.4	4.7	8.9	14.2	12.9	58.4	**72.6**
Average	37.0	49.2	49.0	52.7	44.7	50.5	44.4	**57.0**
RefCOCO								
RefCOCOm	54.0	64.4	64.7	72.1	62.0	**87.1**	36.4	77.3

Impact of resolution: We analyze the effect of input resolution on dense perception performance by evaluating Falcon Perception across a range of resolutions from 448^2 to 1024^2. As shown in Figure 5, increasing resolution yields consistent performance gains, but the impact is non-uniform across metrics and tasks. We observe a "Dense" phase transition on **PBench Dense stress test**, where the model is effectively blind at 448^2 (3.9% micro-F1) but reaches 61.0% at 1024^2. This **15**×

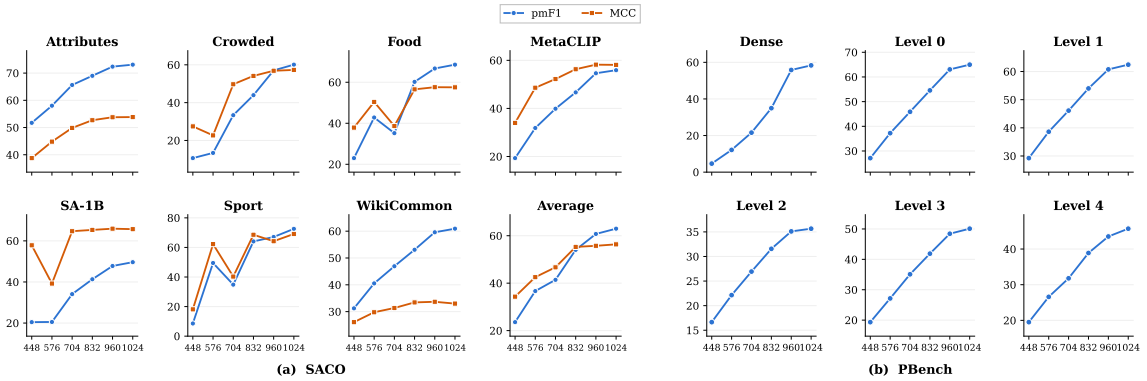
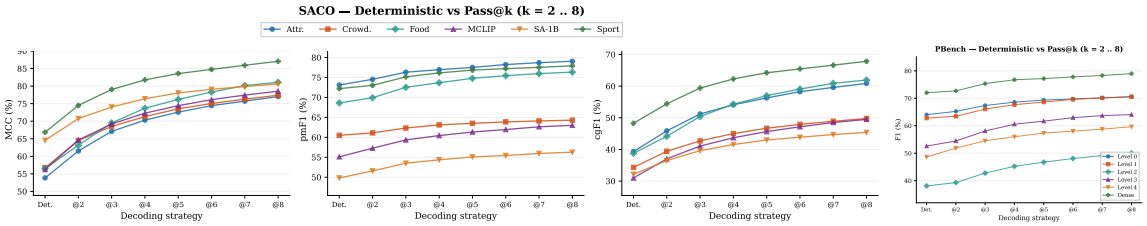


Fig. 5: Resolution Scaling. Performance as a function of image resolution. While semantic tasks (*e.g.*, Attributes) are relatively robust to lower resolutions, dense and small-object tasks (*e.g.*, Crowded, Sports, Dense) exhibit a phase transition, requiring high resolution to resolve individual instances.



(a) SACO **(b)** PBench

Fig. 6: Pass@k Performance scaling. Evolution of detection metrics across SACO (left) and PBench (right) as the number of samples k increases. Performance consistently scales with k, particularly on the most challenging subsets.

improvement confirms that for crowded scenes, while the transformer backbone can understand semantics at low resolution, the "bottleneck" for dense perception is strictly spatial: resolving a high number of details in the input image is a prerequisite for the specialized heads to operate effectively. Furthermore, we notice a decoupling between recognition and grounding. At 448^2, the model achieves a respectable IL_MCC of 0.34 (for detecting presence) but a poor pmF1 of 23.6 (for drawing masks). As resolution increases to 1024^2, localization quality (pmF1) improves by $\mathbf{2.7}\times$, while classification (MCC) grows more modestly by $\mathbf{1.6}\times$. This suggests that low-resolution features are sufficient for semantic recognition ("Is there a cat?"), but high-resolution input is necessary for precise spatial grounding ("where exactly is the cat?"), highlighting the importance of our backbone design that can handle image at arbitrary resolution and native aspect ratio. More details are given in supplementary (Section D).

6.2 Effect of Sampling

To further investigate the latent capabilities of Falcon Perception, we evaluate its performance under a sampling-based inference regime, and report Pass@k (best score out of k predictions). Doing so, we hypothesize that the maximum likelihood objective used during training allows the model to

capture a rich distribution of potential outputs, some of which may be more accurate than the most certain prediction in complex scenarios.

Sampling Mechanism: We enable stochastic decoding by sampling from the output heads of the model across three primary dimensions:

- *Language Head:* We sample from the vocabulary logits, allowing for diversity in the generated tokens. This mainly influences the end of generation, *i.e.*, whether the model decides whether there is an object corresponding to the query in the image and how many objects to detect and segment.
- *Center Head:* We sample from the x and y bin logits, which govern the localization of object centers. Thus, this influences the exact localization of the object and which object to detect at the current timestep.
- *Size Head:* We sample from the h and w bin logits, influencing the predicted dimensions of the bounding boxes.

Experimental Protocol: We evaluate the model on both the SA-Co and PBench benchmarks. For each test instance, we generate k independent samples and select the "best" prediction (Pass@k) based on the task-specific metric. We compare the baseline deterministic (argmax) performance against $k \in \{2, 4, 6, 8\}$.

Results: As shown in Figure 6, performance across all metrics and subsets consistently and substantially improves as k increases. In particular, we note that we outperform the deterministic baseline directly from pass@2 (details in supplementary Table A.2).

On the SA-Co benchmark, the average cgF_1 score jumps from 34.7 (Baseline) to 54.3 (Pass@8), on par with SAM3, *i.e.*, a **+19.6** point absolute gain. The most significant delta is observed in the *Wiki-Common* subset, where cgF_1 improves from 19.3 to 45.0, outperforming SAM3 from pass@4.

On PBench, we observe a similar trend across all difficulty levels. The "hard" scenarios show the most significant recovery: for *Level 2*, *Level 3*, and *Level 4*, F_1 scores increase by **+12.0**, **+11.5**, and **+11.0** points respectively when moving from deterministic to Pass@8. Furthermore, we find that sampling is particularly effective in "hard" scenes with extreme occlusion, dense object clusters, or expressions requiring complex reasoning. In these instances, the model's first-best prediction might fail, but the underlying probability distribution often contains the correct solution within the top few samples (see supplementary Figure A.8 for the qualitative results).

6.3 OCR Extension

Pipeline: We follow a two-stage approach used by recent OCR systems: (i) an off-the-shelf layout detector proposes element regions (text blocks, tables, formulas, figures), and (ii) Falcon-OCR performs element-level recognition on each cropped region, conditioned on the element type. Text is emitted as plain text, formulas as LaTeX, and tables as HTML; outputs are assembled into a structured document by reading order.

Model and training: Falcon-OCR uses the same early-fusion dense transformer backbone as Falcon Perception (shared patch/text tokenization and hybrid attention masking), but is trained *from scratch* to learn OCR-specific visual features, without vision distillation. We train on a mixture of document parsing, formula recognition, and table structure data (public + synthetic rendered sources). The objective is standard next-token prediction over the structured output sequences.

Benchmarks: We evaluate on **olmOCR** (diverse real-world documents; binary correctness checks) and **OmniDocBench** (full-page parsing with continuous metrics: Edit Distance↓, CDM↑ for

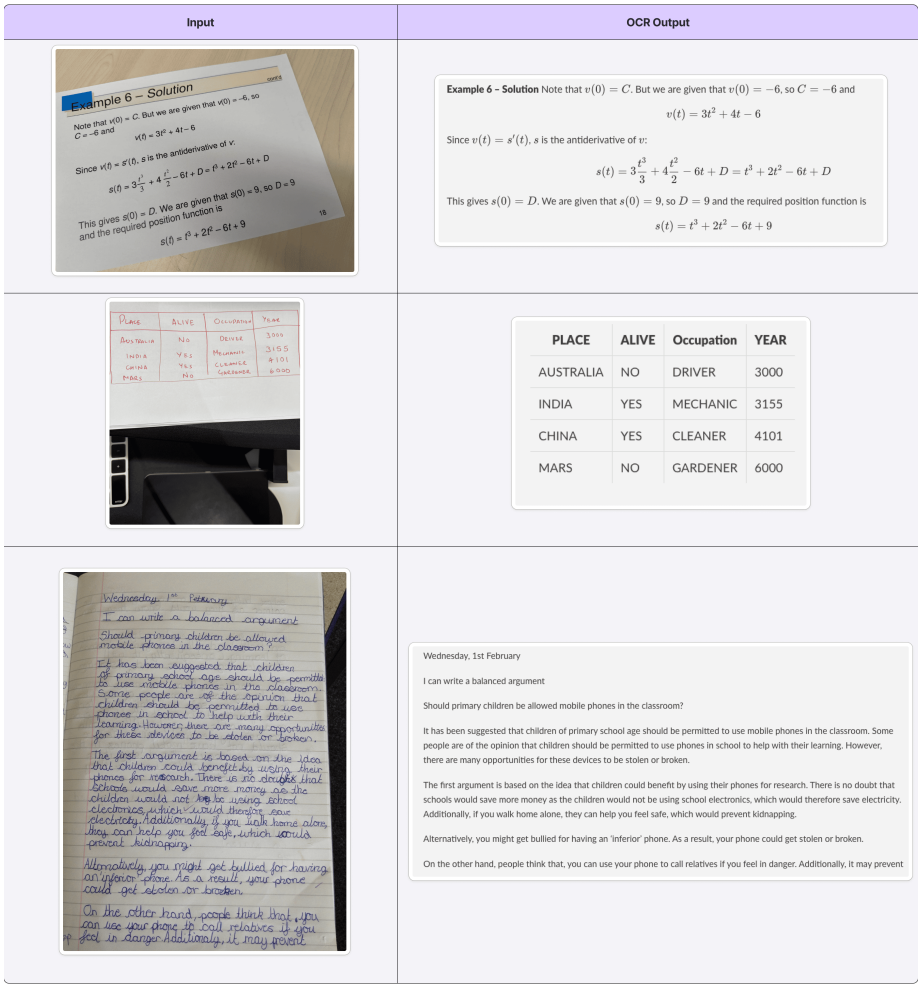


Fig. 7: Example OCR output for inputs comprising formulae, tables and handwriting (from top to bottom). Best viewed zoomed-in.

formulas, and TEDS↑ for tables). As shown in Table 7, Falcon-OCR is competitive despite its compact size, with strong performance on multi-column layouts and tables as shown in Figure 7. The remaining failures are dominated by tiny text and heavily degraded scans; detailed breakdowns and qualitative examples are provided in the supplementary material (Section E).

6.4 Initialization: Random *vs.* Distillation

Here, we analyze the impact of initializing the Falcon-Perception weights using the multi-teacher distillation, detailed in supplementary Section A.1. To this end, we pretrain and decay (stages 1 and 2 of our recipe) a smaller 300M-sized model on the detection task (bounding boxes) with different weights initialization. From Figure 8, we observe that even after 230K steps of training across both

Table 7: Average performance on olmOCR and OmniDocBench.

Model	olmOCR Avg (%)	OmniDocBench Overall↑
Mistral OCR 3 [39]	81.7	85.20
Chandra [14]	**82.0**	88.97
PaddleOCR VL 1.5 [13]	79.3	**94.37**
DeepSeek OCR v2 [53]	78.8	87.66
Ours: Falcon-OCR (300M)	80.3	88.64

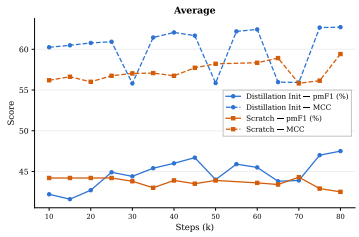


Fig. 8: Impact of initialization: Random initialization results in lower performance, compared to distillation initialization, even after being trained long enough for 230K steps over pretraining and decay (stages 1 and 2) combined.

stages, the average performance of randomly initialized model lags the distillation initialized model on the Sa-Co benchmark on both pmF1 and classification (MCC) metrics by a significant margin. Additionally, we note that training with scratch initialization for the segmentation task diverged in the early stages of pretraining itself. This is likely attributed to the fact that accurately predicting the instance masks requires enhanced image features that encode the structure of the scene, which is not possible in early stages of the training with random initialization. Consequently, it results in high gradients for image tokens leading to divergence of the model weights. Both these ablations clearly indicate the necessity and effectiveness of the multi-teacher distillation stage in our training pipeline.

7 Conclusion

Falcon Perception revisits a common assumption in dense prediction that perception requires a separate vision encoder and a dedicated decoder and shows that, for dense grounding tasks, a single early-fusion Transformer can serve as both. By processing image patches, text, and task tokens in one shared parameter space with a hybrid attention pattern, Falcon Perception removes encoder–decoder cross-attention bottlenecks while retaining an autoregressive interface that naturally supports native image size and aspect ratio, flexible text query and variable numbers of instances. Empirically, Falcon Perception achieves strong open-vocabulary segmentation performance, with gains that grow as prompts become more compositional and as scenes become more crowded, while resolution scaling reveals a clear "dense" phase transition that highlights spatial detail, not recognition, as the dominant bottleneck in challenging settings. We view these results through a "bitter lesson" lens: rather than addressing each failure mode with new modules, Falcon Perception keeps a single scalable backbone and shifts progress toward data, compute, and training signals, using lightweight heads only where outputs are continuous and dense. Finally, we demonstrate that the same unified backbone extends beyond segmentation to document OCR: a compact Falcon-OCR model with only 300M parameters attains competitive performance on standard OCR benchmarks.

References

1. Adept.ai: Fuyu-8b: A multimodal architecture for ai agents (2026), https://www.adept.ai/blog/fuyu-8b/, accessed 2026-03-03 2

2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv: 2308.12966 (2023). https://doi.org/10.48550/arXiv.2308.12966 2
3. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D.M., Neumann, M., Alabdulmohsin, I.M., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J.M., Kabra, R., Bauer, M., Bovsnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai, X.Q.: Paligemma: A versatile 3b vlm for transfer. ArXiv **abs/2407.07726** (2024). https://doi.org/10.48550/arXiv.2407.07726 6
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts. arXiv preprint (2025). https://doi.org/10.48550/arXiv.2511.16719 1, 2, 10, 11
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. European Conference on Computer Vision (2020). https://doi.org/10.1007/978-3-030-58452-8_13 1, 2
6. Chaybouti, S., Narayan, S., Dahou, Y., Khắc, P.H.L., Singh, A., Huynh, N.D., Para, W.R., Kuehne, H., Hacid, H.: Siglino: Efficient multi-teacher distillation for agglomerative vision foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026) 7, 8
7. Chen, J., Guo, L., Sun, J., Shao, S., Yuan, Z., Lin, L., Zhang, D.: EVE: efficient vision-language pre-training with masked prediction and modality-aware moe. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 1110–1119. AAAI Press (2024). https://doi.org/10.1609/AAAI.V38I2.27872 3
8. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv: 2109.10852 (2021). https://doi.org/10.48550/arXiv.2109.10852 1, 2
9. Chen, T., Saxena, S., Li, L., Lin, T.Y., Fleet, D.J., Hinton, G.E.: A unified sequence interface for vision tasks. Advances in Neural Information Processing Systems **35**, 31333–31346 (2022) 1, 2
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv: 2312.14238 (2023). https://doi.org/10.48550/arXiv.2312.14238 2
11. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 6
12. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint (2025). https://doi.org/10.48550/arXiv.2507.06261 11
13. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., et al.: Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. arXiv preprint arXiv:2601.21957 (2026). https://doi.org/10.48550/arXiv.2601.21957 15
14. Datalab: Chandra: Layout-aware ocr for structured document understanding. https://github.com/datalab-to/chandra (2025), accessed: 2026-03-05 15
15. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14987–14997 (2025) 11

16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv: 1311.2524 (2013). https://doi.org/10.48550/arXiv.1311.2524 2
17. Huh, M., Cheung, B., Wang, T., Isola, P.: Position: The platonic representation hypothesis. In: Forty-first International Conference on Machine Learning (2024) 2
18. Huynh, N.D., Le-Khac, P.H., Para, W.R., Singh, A., Narayan, S.: Vision-language models can't see the obvious. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24159–24169 (October 2025) 2
19. Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/, accessed: 2026-06-27 8
20. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr - modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021) 1
21. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014) 3
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 1, 10
23. Kolesnikov, A., Susano Pinto, A., Beyer, L., Zhai, X., Harmsen, J., Houlsby, N.: Uvim: A unified modeling approach for vision with learned guiding codes. Advances in Neural Information Processing Systems **35**, 26295–26308 (2022) 1
24. Korrapati, V.: Moondream2. Hugging Face Model Hub (2025). https://doi.org/10.57967/hf/6762, https://huggingface.co/vikhyatk/moondream2, accessed: 2026-06-27 2, 10, 11
25. Lei, W., Wang, J., Wang, H., Li, X., Liew, J.H., Feng, J., Huang, Z.: The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20758–20769 (2025) 3
26. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020). https://doi.org/10.48550/arXiv.2006.16668 3
27. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10965–10975 (June 2022) 1
28. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model. arXiv preprint arXiv: 2512.02498 (2025). https://doi.org/10.48550/arXiv.2512.02498 2
29. Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., Yih, W.t., Zettlemoyer, L., et al.: Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. arXiv preprint arXiv:2411.04996 (2024). https://doi.org/10.48550/arXiv.2411.04996 3
30. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Neural Information Processing Systems (2022). https://doi.org/10.48550/arXiv.2203.02053 2
31. Lin, X.V., Shrivastava, A., Luo, L., Iyer, S., Lewis, M., Gosh, G., Zettlemoyer, L.S., Aghajanyan, A.: Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. ArXiv **abs/2407.21770** (2024). https://doi.org/10.48550/arXiv.2407.21770 3
32. Liu, J., Su, J., Yao, X., Jiang, Z., Lai, G., Du, Y., Qin, Y., Xu, W., Lu, E., Yan, J., Chen, Y., Zheng, H., Liu, Y., Liu, S., Yin, B., He, W., Zhu, H., Wang, Y., Wang, J., Dong, M., Zhang, Z., Kang, Y., Zhang, H., Xu, X., Zhang, Y., Wu, Y., Zhou, X., Yang, Z.: Muon is scalable for llm training. arXiv preprint arXiv: 2502.16982 (2025). https://doi.org/10.48550/arXiv.2502.16982 8

33. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023). https://doi.org/10.48550/arXiv.2303.05499 1, 11
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 8
35. Luo, G., Yang, X., Dou, W., Wang, Z., Liu, J., Dai, J., Qiao, Y., Zhu, X.: Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24960–24971 (2025) 3
36. Maniparambil, M., Akshulakov, R., Djilali, Y.A.D., El Amine Seddik, M., Narayan, S., Mangalam, K., O'Connor, N.E.: Do vision and language encoders represent the world similarly? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14334–14343 (2024) 2
37. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. ArXiv **abs/2205.06230** (2022). https://doi.org/10.48550/arXiv.2205.06230 2
38. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2023). https://doi.org/10.48550/arXiv.2306.09683 1, 11
39. Mistral AI: Mistral ocr 3. https://mistral.ai/news/mistral-ocr-3 (2025), accessed: 2026-03-05 15
40. Moondream: Moondream 3 (preview). Hugging Face Model Hub (2025). https://doi.org/10.57967/hf/6761, https://huggingface.co/moondream/moondream3-preview, accessed: 2026-06-27 2, 10, 11
41. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. Advances in Neural Information Processing Systems **35**, 9564–9576 (2022) 2
42. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24838–24848 (2025) 2
43. Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., Wolf, T., et al.: The fineweb datasets: Decanting the web for the finest text data at scale. arXiv preprint arXiv:2406.17557 (2024). https://doi.org/10.48550/arXiv.2406.17557 7
44. Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models. arXiv preprint arXiv:2502.18443 (2025). https://doi.org/10.48550/arXiv.2502.18443 2
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv: 2103.00020 (2021). https://doi.org/10.48550/arXiv.2103.00020 2
46. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR) (2025). https://doi.org/10.48550/arXiv.2408.00714 1
47. Ren, T., Chen, Y., Jiang, Q., Zeng, Z., Xiong, Y., Liu, W., Ma, Z., Shen, J., Gao, Y., Jiang, X., et al.: Dino-x: A unified vision model for open-world object detection and understanding. arXiv preprint arXiv:2411.14347 (2024). https://doi.org/10.48550/arXiv.2411.14347 11
48. Shen, Y., Fu, C., Chen, P., Zhang, M., Li, K., Sun, X., Wu, Y., Lin, S., Ji, R.: Aligning and prompting everything all at once for universal visual perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13193–13203 (2024) 11
49. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025). https://doi.org/10.48550/arXiv.2508.10104 7, 8

50. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020) 6
51. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025). https://doi.org/10.48550/arXiv.2502.14786 7
52. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. arXiv preprint arXiv: 2505.23941 (2025). https://doi.org/10.48550/arXiv.2505.23941 2
53. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr 2: Visual causal flow. arXiv preprint arXiv:2601.20552 (2026). https://doi.org/10.48550/arXiv.2601.20552 15
54. Wimmer, T., Truong, P., Rakotosaona, M.J., Oechsle, M., Tombari, F., Schiele, B., Lenssen, J.E.: Anyup: Universal feature upsampling. arXiv preprint arXiv:2510.12764 (2025). https://doi.org/10.48550/arXiv.2510.12764 3, 6
55. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4818–4829 (2024) 2
56. Xiong, J.: On n-dimensional rotary positional embeddings (2025), https://jerryxio.ng/posts/nd-rope/, accessed: 2026-06-30 7
57. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025). https://doi.org/10.48550/arXiv.2505.09388 2, 10, 11
58. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016) 3
59. Zhang, T., Li, X., Huang, Z., Li, Y., Lei, W., Deng, X., Chen, S., Ji, S., Feng, J.: Pixel-sail: Single transformer for pixel-grounded understanding. arXiv preprint arXiv:2504.10465 (2025). https://doi.org/10.48550/arXiv.2504.10465 3