This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Text-guided Visual Prompt DINO for Generic Segmentation

Yuchen Guan[1*‡], Chong Sun[2*†], Canmiao Fu[2*], Zhipeng Huang[2], Chun Yuan[1†], Chen Li[2]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]WeChat AI, Tencent Inc.



Figure 1. **Detection/Segmentation samples.** Our Prompt-DINO is a text-guided visual Prompt DINO. Prompt-DINO supports single or multiple visual or text prompts for in-image or cross-image generic segmentation. Based on the granularity of the provided prompts, Prompt-DINO is capable of performing both panoptic and part segmentation.

Abstract

Recent advancements in multimodal vision models have highlighted limitations in late-stage feature fusion and suboptimal query selection for hybrid prompts open-world segmentation, alongside constraints from caption-derived vocabularies. To address these challenges, we propose Prompt-DINO, a text-guided visual Prompt DINO framework featuring three key innovations. First, we introduce an early fusion mechanism that unifies text/visual prompts and backbone features at the initial encoding stage, enabling deeper cross-modal interactions to resolve semantic ambiguities. Second, we design order-aligned query selection for DETR-based architectures, explicitly optimizing the structural alignment between text and visual queries during decoding to enhance semantic-spatial consistency. Third, we develop a generative data engine powered by the Recognize Anything via Prompting (RAP) model, which synthesizes 0.5B diverse training instances through a dual-path cross-verification pipeline, reducing label noise by 80.5% compared to conventional approaches. Extensive experiments demonstrate that Prompt-DINO achieves state-of-the-art performance on open-world detection bench-

* Equal Contribution.
† Corresponding Authors. Emails: waynecsun@tencent.com, yuanc@sz.tsinghua.edu.cn
‡ Work done as interns at WeChat.

marks while significantly expanding semantic coverage beyond fixed-vocabulary constraints. Our work establishes a new paradigm for scalable multimodal detection and data generation in open-world scenarios. Data&Code are available at https://github.com/WeChatCV/WeVisionOne.

1. Introduction

Recent advancements in vision models [5, 43–46, 58] have focused on unifying diverse prompting strategies—such as text, visual, or hybrid inputs—to enable versatile object understanding and segmentation. Pioneering works like [16, 21, 22, 60] have demonstrated the potential of multimodal prompting for tasks ranging from open-set detection to segmentation. Despite their emerging capabilities, open-world detection and segmentation methods leveraging multimodal prompts demonstrate inherent technical deficiencies.1) *Late-stage fusion of multimodal features*. [22, 60] process text, visual or other prompts, and backbone features in separate subspaces, merging them only in later encoding stages. [16] similarly adopts contrastive learning to align modalities without joint optimization at the encoding stage. This delayed interaction restricts cross-modal alignment, particularly in ambiguous scenarios where early fusion of text and visual cues could disambiguate object semantics. 2) *Insufficient exploration of query selection mechanisms*. Most of these methods employ generic cross-attention for text-guided query selection, or handle queries corresponding to text and visual prompts independently, none of which explicitly optimizes the structural alignment between queries corresponding to text and visual cues respectively. Like [16] processes text and visual prompts independently. This results in the selection of sub-optimal queries that do not promote optimization between text and images.

Beyond these architectural limitations, a broader challenge persists across the field: the reliance on caption-derived vocabularies or fixed-vocabulary, which inherently constrain generalization to novel or composite objects. [16, 22, 60] derive their object vocabularies from entity words extracted from image captions or use fixed-vocabulary detector(like TAP [32]). While this approach avoids manually predefined categories, it remains bounded by the entities explicitly mentioned in the training captions. For instance, TRex-2 [16] employ TAP [32] to annotate SA-1B [19] with a category name from a dictionary of 2560 classes. Nevertheless, the reality that SA-1B [19] is associated with a vocabulary significantly exceeding 2560 substantially constrains the dataset's potential.

In this work, we propose Prompt-DINO, a novel text-guided visual Prompt DINO, that addresses these challenges through two architectural innovations and a paradigm-shifting data engine. We pioneer an *early fusion mechanism* that deeply integrates text, visual, and backbone features at the initial encoding stage. Unlike [16, 22, 60], our model unifies multimodal signals into a cohesive representation from the outset. This enables richer cross-modal interactions, reducing ambiguity in open-set scenarios. And we introduce *order-aligned query selection*, a novel paradigm for DETR-based architectures that explicitly optimizes the alignment between text queries and visual queries during decoding. By enhancing semantic correlation and spatial consistency within query priorities, our method facilitates the optimization of text and image query selection in a unified direction, thereby surpassing the heuristic cross-attention mechanisms employed in prior research. Also, we overcome the fixed vocabulary bottleneck through a generative data engine with *dual-path cross-verification pipeline*. We proposal an novel *Recognize Anything via Prompting (RAP)* model, a 0.5B-parameter multimodal large language model(MLLM) that generates semantic labels for input images and their corresponding masks. Our engine leverages RAP to synthesize 0.5B diverse training instances with enhanced semantic relevance. To ensure label fidelity, we further implement a novel dual-path cross-verification pipeline that substantially reduces label noise by 80.5% compared to conventional single-path approaches. This integrated data engine pioneers scalable open-world dataset generation, bridging limited annotation vocabularies with real-world detection needs.

To summarize, our contributions are threefold:

- We propose an early fusion mechanism that integrates text, visual, and backbone features during initial encoding stage, unifying multimodal signals into a cohesive representation.
- We introduce order-aligned query selection, a new paradigm for DETR-based architectures that explicitly optimizes the alignment between text queries and visual queries during decoding.
- We develop a generative data engine powered by our RAP, incorporating a dual-path cross-verification pipeline. This approach synthesizes diverse and semantically relevant training data, and establishes a new paradigm for scalable open-world data generation.

2. Related Work

2.1. General Recognition via Prompts

Recent advances in general recognition models (e.g., detection/segmentation [10, 11, 18, 24, 27, 30, 49, 53, 56]) have demonstrated that multimodal alignment between vision and language modalities serves as the cornerstone for open-world understanding. These frameworks typically employ joint text-visual prompting mechanisms to achieve object localization through aligned cross-modal feature spaces.

Text Prompt Paradigm. Existing methods predominantly focus on text-based cross-modal alignment. GLIP

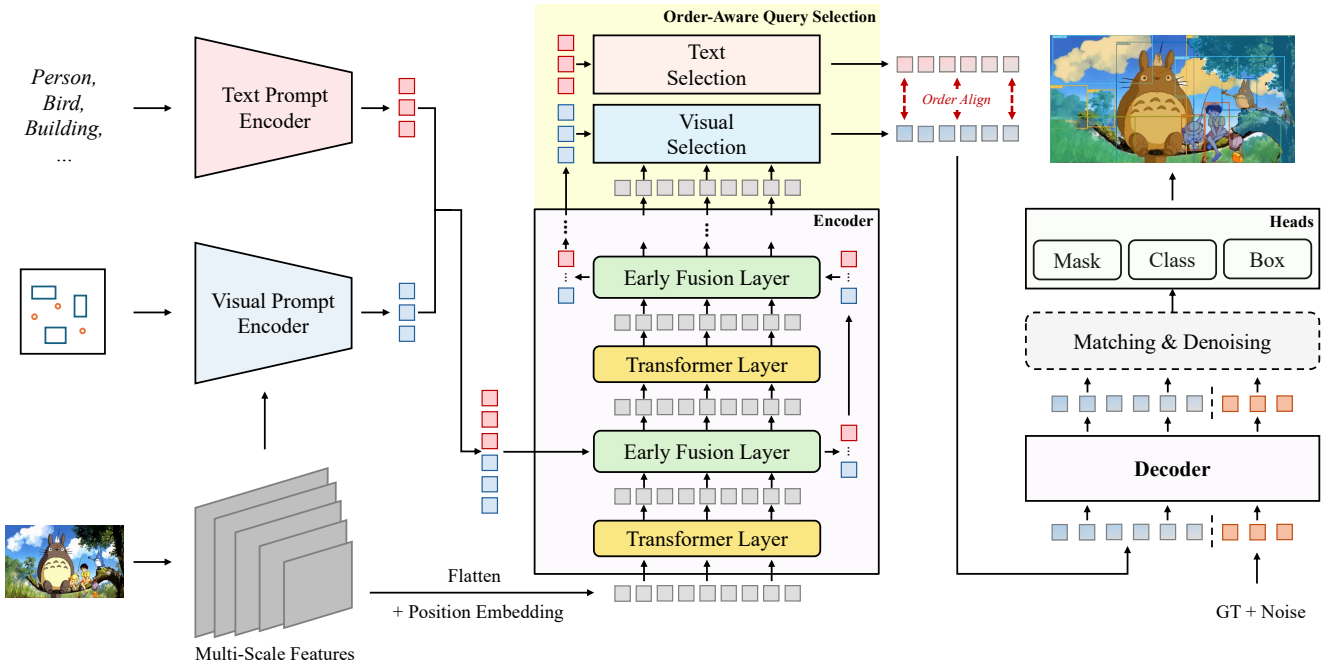


Figure 2. **The Overall Framework of Prompt-DINO.** Prompt-DINO is constructed based on the DINO architecture. Images are processed by ViT to extract multi-scale features, while visual and text prompts are processed by deformable attention transformer layers and CLIP's text encoder, respectively, to derive embeddings. Through early fusion layers in the backbone encoder and perception-aware query selection, Prompt-DINO perform deep fusion and alignment of visual and text in two steps, with the final predictions generated by the Decoder.

[24] pioneers the adaptation of image-text pretraining to downstream detection tasks through deep feature fusion, establishing language-aware visual representations. GroundingDINO [27] further extends this paradigm by demonstrating that multi-stage cross-attention fusion significantly enhances performance in open-vocabulary detection. While these approaches achieve remarkable progress, their heavy reliance on textual semantics limits their capacity to handle visual concepts with ambiguous linguistic descriptions.

Visual Prompt Paradigm. Complementary to text prompts, visual prompting (points/boxes/masks) provides geometric guidance for specific object localization. SAM [19] establishes a milestone in interactive segmentation through decoupled visual prompt encoding, though lacking semantic awareness. Subsequent works like SEEM [60] and DINOv [22] enhance this paradigm by integrating semantic context and exemplar-based reasoning. A notable hybrid approach is T-Rex2 [16], which combines both modalities but still treats visual prompts as auxiliary geometric cues. Our key insight lies in that existing methods inadequately exploit the semantic synergy between visual prompts and language guidance, leaving visual prompting's full potential underdeveloped for language-aware feature learning.

2.2. Data Engine for Open-World Learning

Scalable data curation has become pivotal for training robust open-world detector/segmentation models. SAM2 [36] introduces a cyclic model-annotator collaboration framework, iteratively refining the SA-V dataset through human-machine interactions. Alternative strategies employ cross-model verification: YOLO-World [6] curates CC3M-Lite by relabeling CC3M [39] with GLIP and CLIP, while T-Rex2 [16] propagates pseudo-labels across SA-1B and LAION400M [34]. Current methods face three key limitations: (1) Over-reliance on noun phrase extraction from captions misses more detailed concepts and introduces too many redundant nouns; (2) Single-path verification propagates annotation errors; (3) Semantic ambiguity in free-form text introduces label noise. Our data engine addresses these issues and enables the automatic generation of high-quality training data with reduced noise.

3. Method

Motivated by the limitations observed in existing hybrid prompting methods for detection and segmentation tasks, we introduce Prompt-DINO, an end-to-end generic segmentation model built upon the DINO architecture [52]. Our model natively integrates multimodal prompting capabilities for both textual and visual inputs while maintaining architectural simplicity. In this section, we first formalize the promptable segmentation paradigm and outline the core design prompt encoders of Prompt-DINO (§3.1). We then systematically address three critical challenges: (1) early fusion with visual prompts through cross-modal feature align-

ment(§3.2), (2) order-aware query selection for maintaining semantic consistency(§3.3) and (3) construction of a scalable data engine for open-world generalization(§3.4). Finally, we introduce the training objective(§3.5).

3.1. Promptable Segmentation Task

The Prompt-DINO processes an input image $\mathcal{I}$, with accompanying visual prompts $\mathcal{P}^v = \{\mathcal{P}_1^v, ..., \mathcal{P}_K^v\}$, and textual prompts $\mathcal{P}^t = \{\mathcal{P}_1^t, ..., \mathcal{P}_K^t\}$, where K is the number of interactive prompts per image. Following established architectures in DETR [57] and DINO [52], our system first extracts hierarchical visual features $F = \{f_1, ..., f_L\}$ through a vision backbone network (e.g., ViT [9] or Swin Transformer [28]), where L represents the number of feature pyramid levels.

Visual Prompt Encoder. Specifically, our visual encoding module employs L transformer blocks with deformable attention mechanisms [22]. For each target object, we initialize a learnable query embedding q_i and progressively refine it through multi-scale feature aggregation. The transformation at each layer l can be formulated as:

$$v_i^l = \textit{MSDA}(v_i^{l-1}, f_l), l \in [1, ..., L], \quad \text{with} \quad v_i^0 = q_i \quad (1)$$

where $\textit{MSDA}(\cdot)$ denotes multi-scale deformable attention operation. The final visual prompt representations $V = \{v_1, ..., v_K\}$ are obtained as $v_i = v_i^L$.

Text Prompt Encoder. Building upon insights from recent open-world detection studies, we employ a CLIP [34] text encoder rather than conventional BERT [8] approaches used in earlier models like [24, 27]. This design choice enables better remantic correlation modeling for input words, phrases, and short sentences, generating text prompts $T = \{t_1, ..., t_K\}$ that effectively guide cross-modal feature alignment.

The integrated representation (text(T), visual(V), multi-scale features(F)) feeds into the backbone encoder for multimodal fusion, enabling dynamic adaptation to prompts while preserving cross-modal alignment.

3.2. Early Fusion with Visual Prompts

We are the first to propose the early fusion framework that integrates text prompts, visual prompts, and input image features through multi-modal interaction. While this fusion mechanism effectively reduces the modality gap between textual and visual features, enhances prompt-image correlation, and improves target recall, We found unintended consequences of severe hallucinations in our experiments. This phenomenon stems from the inherent limitation of conventional cross-attention mechanisms: they compulsively reconstruct Key features most similar to the Query features, even when all Key features are substantially dissimilar to the current Query. To address this non-target feature reconstruction issue, we introduce a novel gated cross-attention mechanism incorporating a learnable background token (B). Reconstruction occurs exclusively when the Query-Key similarity exceeds an adaptive threshold; otherwise, the background token automatically activates to suppress hallucination. As illustrated in Fig. 3, our architecture comprises three core components: 1) a self-attention layer for intra-modal refinement, 2) the proposed gated cross-attention layer for inter-modal interaction, and 3) a feedforward neural network (FFN) for feature transformation. The cross-attention module implements three interaction pathways: feature-to-visual prompt, visual prompt-to-feature, and feature-to-text prompt. For visualization clarity, Fig. 3 presents only stage-2 features mapped to text space through our two-phase training scheme (detailed in §3.5). The enhanced cross-modal fusion operates through the following formulation:

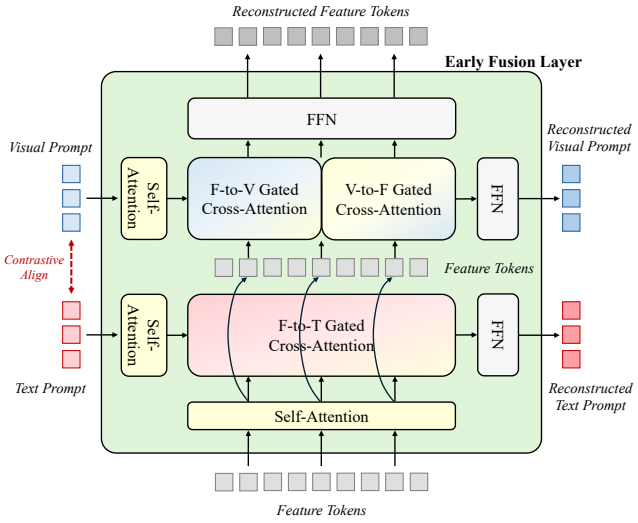


Figure 3. **Structure of Early Fusion Layer.** Notably, the contrastive alignment of visual and text prompts is completed before early fusion. By employing Gated Cross-Attention, text prompts are reconstructed with image features, while visual prompts and image features mutually reconstruct each other.

$$\begin{aligned} t_i' &= \text{GatedAttn}(f_i, t_i, t_i) \\ v_i' &= \text{GatedAttn}(f_i, v_i, v_i) \\ f_i' &= \text{GatedAttn}(v_i, f_i, f_i) \end{aligned} \quad (2)$$

where the gated attention mechanism is defined as:

$$\text{GatedAttn}(Q, K, V) = \text{softmax}(\tfrac{Q \cdot [K, B]}{\sqrt{d_k}})[V, B] \quad (3)$$

Here, B denotes our learnable background token. Following TRex-2 [16] methodology, we employ region-level contrastive learning to align visual and textual modalities, utilizing symmetric cross-entropy loss [34] for embedding alignment:

$$\mathcal{L}_{align} = -\frac{1}{K} \sum_{i=1}^{K} \log \frac{v_i \cdot t_i}{\sum_{j=1}^{K} v_i \cdot t_j} \quad (4)$$

In our experience, symmetric cross-entropy losses have better convergence than InfoNCE loss [31].

In the context of training detection and segmentation tasks, it is common practice to construct negative samples. For negative prompt construction, while text prompts naturally leverage category vocabularies for negative sampling, visual prompts require explicit negative construction. We address this by aggregating visual prompt features across all categories within each batch and computing inter-category averages to generate negative visual prompts. This approach ensures sufficient diversity in negative samples while maintaining semantic relevance.

To mitigate annotation granularity discrepancies across datasets (e.g., "bird" in Dataset A vs. "pigeon" in Dataset B), we implement a dataset-aware sampling strategy. Our novel data sampler ensures each training batch contains only intra-dataset samples, preventing semantically similar but categorically distinct concepts from becoming conflicting negative samples. This is particularly critical in visual-prompt open-world detection/segmentation scenarios where label hierarchies may differ substantially across source datasets.

3.3. Order-aware query selection

Prompt-DINO adopts a DETR-based architecture that selects top-K encoder features through similarity computation between encoder features and prompt embeddings, following DINO [52] and MaskDINO [21]. However, our approach fundamentally diverges from existing methods by simultaneously considering both textual-visual query similarity and inter-query correlations with visual prompts. To further strengthen the alignment between visual and textual prompts, we introduce order-aware alignment during query selection, which establishes the second-level alignment between visual and textual prompts during training. Specifically, we employ Kendall's τ coefficient to quantify the ordinal consistency [12] between queries selected by visual prompts and text prompts, the Kendall's τ coefficient can be expressed as:

$$\tau = \frac{P_c - P_d}{\frac{1}{2}N(N-1)} \tag{5}$$

where P_c represents concordant pairs and P_d represents discordant pairs, N represents the total number of queries. Given the similarity score between query embeddings and prompt embeddings $S^v = s_1^v, ..., s_N^v$ and $S^t = s_1^t, ..., s_N^t$, we calculate P_c and P_d as follows:

$$\tau = \frac{\sum_i \sum_{j<i} \operatorname{sgn}(s_i^t - s_j^t) \cdot \operatorname{sgn}(s_i^v - s_j^v)}{\frac{1}{2}N(N-1)} \tag{6}$$

Since the sgn is not differentiable, we approximate the sign function with the $tanh$ function and utilize the opposite of τ as the order alignment loss:

$$\mathcal{L}_{order} = -\frac{\sum_i \sum_{j<i} \tanh(s_i^t - s_j^t) \cdot \tanh(s_i^v - s_j^v)}{\frac{1}{2}N(N-1)} \tag{7}$$

Our proposed alignment mechanism operates at two complementary levels: The primary alignment directly optimizes the learning objective through $\mathcal{L}_{align}$ (as given in Eq. 4), while our novel order-aware alignment provides indirect assurance through two critical aspects: (1) Semantic proximity between corresponding text and visual prompts for candidate queries naturally reduces overall training complexity by minimizing feature space discrepancies; (2) Training-inference consistency preservation ensures equivalent substitutability between text and visual prompts during inference through maintained representational parity.

Fig. 5 demonstrates the effectiveness of our proposed method. Prior to applying the order-aware loss optimization, we observe significant disorder in both textual and visual prompt queries, with no discernible alignment between corresponding semantic representations. The optimization process induces a remarkable transformation - post-optimization analysis reveals substantially enhanced consistency between semantically equivalent textual and visual prompt queries.

3.4. Data Engine

Effective visual concept learning from rich image-text pairs is fundamental to enhancing the generalization capabilities of open-vocabulary object detectors. Existing methods predominantly rely on captioning models([3, 23, 26] etc.), coupled with lexical analysis tools like SpaCy [14] to extract noun phrases. These noun phrases are subsequently employed to generate pseudo-labels via open-set detection models([24, 27]). Nonetheless, these techniques have significant drawbacks. They frequently omit valid labels, introduce erroneous annotations, and are susceptible to irrelevant lexical noise. All of these issues detrimentally affect detection performance.

To address these challenges, we propose a novel *Recognize Anything via Prompting (RAP)* model, a 0.5B MLLM, which significantly enhances annotation accuracy. Our framework further integrates an innovative dual-path cross-validation pipeline that substantially reduces label noise. This comprehensive approach enables the fully automated generation of high-quality open-world detection datasets while preserving the fidelity of annotations.

Recognize Anything via Prompting. Despite the remarkable capabilities of the Segment Anything Model (SAM) [19] in detecting a wide array of objects within images, and the substantial progress achieved in region-level understanding through multimodal models [1, 2, 33, 35, 51], a critical question persists: Can we effectively assign semantic labels to arbitrary mask prompts? To address this challenge, we introduce RAP, a novel framework that bridges visual segmentation with semantic recognition. As illustrated in Fig. 4, our RAP model processes raw image inputs alongside their corresponding object masks to generate precise semantic annotations for the specified regions. The proposed RAP comprises three key components:

1. Feature extraction. We perform masked pooling over

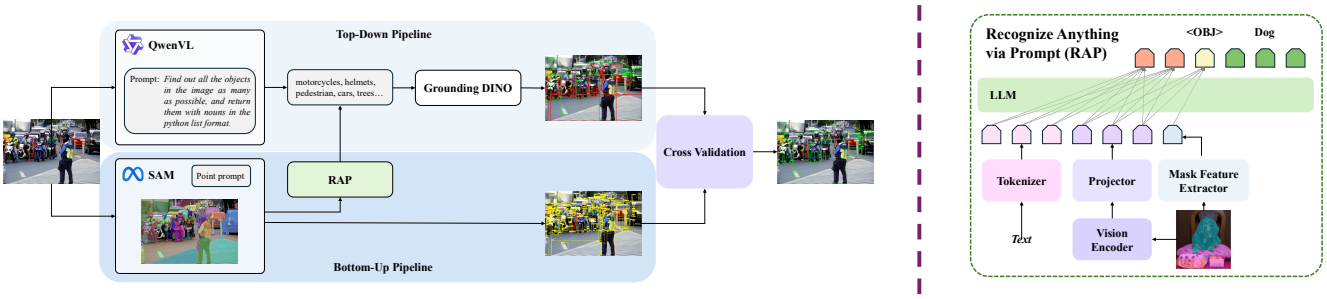


Figure 4. **Data Engine and RAP Frameworks. Left:** The data engine with dual-path cross-verification pipeline. **Right:** The proposed Recognize Anything via Prompting (RAP), a 0.5B-parameter MLLM model.

multi-layer image features to extract region-specific features corresponding to input masks, and concatenate the mask token embeddings obtained via a projection layer with the original image tokens to form a unified representation. This preserves both global contextual information from the complete image and localized object-specific features within masked regions, effectively integrating scene-level understanding with instance-level visual cues.

2.Ground Data Cleaning. To standardize the format of our ground truth data, we leveraged the Qwen model [48] to eliminate redundant descriptive phrases through comprehensive text normalization. Specifically, we transformed verbose expressions into their canonical semantic equivalents (e.g., converting "any part of the sky" to "sky"). This preprocessing strategy serves dual purposes: (1) it reduces the RAP's learning complexity by minimizing lexical variability, and (2) it mitigates the risk of hallucinations by ensuring precise semantic alignment between textual annotations and visual content.

3.Training. The proposed methodology employs a two-stage training strategy. During the first stage, we conduct pre-training using 10M samples, specifically optimizing only the projector while keeping other model components frozen. The second stage implements instruction-tuning, where we unfreeze and jointly optimize the entire model to enhance task-specific adaptation capabilities.

Dual-path cross-verification pipeline. We propose a dual-pipeline annotation framework combining bottom-up and top-down approaches with cross-verification filtering, consisting of three key components:

1. Top-Down Pipeline. Our top-down pipeline leverages the Qwen-VL model [42] to extract semantic tags from images. These tags are subsequently fed into an open-set detection model to generate corresponding bounding boxes. This pipeline demonstrates strong capability in detecting large-scale salient objects (e.g., trees, roads, sky).

2. Bottom-Up Pipeline. To address the inherent limitation of top-down pipeline in capturing fine-grained details, we implement a complementary bottom-up strategy using the RAP integrated with SAM [19]. This combination enables effective detection of smaller objects and detailed components (e.g., eyes, noses, windows) that are frequently missed by conventional top-down pipeline.

3. Cross-Verification. The detection outputs from both pipelines undergo rigorous validation through a three-stage process. First, We employ the Hungarian matching algorithm to establish correspondences between bounding boxes from different pipelines. Next, we utilize the BGE model [47] to compute similarity scores between the tags associated with the matched bounding boxes. Finally, we filter out the candidate pairs whose similarity scores were below the threshold of empirical determination to further clean up the annotated results.

3.5. Training Objective

Building upon previous DINO-like models [52], we employ a combination of L1 loss and GIoU loss [37] for bounding box regression, while for mask segmentation, we adopt cross-entropy loss and dice loss [29]. Following the methodology established in GroudingDINO [27], we utilize contrastive loss between predicted objects and text prompts for classification tasks. To accelerate model convergence, we incorporate denoising training as implemented in DINO [52]. The complete objective function for Prompt DINO can be expressed as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{bbox} + \mathcal{L}_{mask} + \mathcal{L}_{DN} + \mathcal{L}_{align} + \mathcal{L}_{order} \quad (8)$$

Our training methodology adopts a progressive multi-stage strategy: we first optimize the text prompt while keeping the visual components fixed, followed by joint optimization of both textual and visual prompt representations.

4. Experiments

4.1. Settings

In Prompt-DINO, we employ ViT-L [9] as the backbone and fine-tune CLIP-L [34] as our text prompt encoder. The visual prompt encoder consists of 6 layers of deformable cross-attention transformer layers, with a hidden dimension of 256. We adopt the "two-stage" structure proposed by DINO [52], selecting 900 queries for decoding after the backbone encoder.

Method	Backbone	Prompt Type	COCO (in-domain)		LVIS (in-domain)		ADE20K (out-domain)		
			mask AP	box AP	mask AP	box AP	PQ	mask AP	box AP
Mask2Former [4]	Swin-L	Close-Set	48.6	52.1	-	-	-	-	-
OneFormer [15]	Swin-L	Close-Set	48.9	-	-	-	-	-	-
MaskDINO* [21]	Swin-L	Close-Set	50.6	56.2	-	-	-	-	-
kMaX-DeepLab [50]	Swin-L	Close-Set	58.1	-	-	-	-	-	-
GLIPv2 [53]	Swin-H	Text Prompt	48.9	-	-	-	-	-	-
X-Decoder [59]	DaViT-L	Text Prompt	46.7	-	-	-	21.8	13.1	38.1
OpenSeeD [54]	Swin-L	Text Prompt	53.2	58.2	21.0	23.0	19.7	15.0	17.7
T-Rex2 [16]	Swin-L	Text Prompt	-	52.2*	-	45.8*	-	-	-
APE [40]	ViT-L	Text Prompt	49.3	58.3	53.0	59.6	27.2	24.4	29.6
DINOv [22]	Swin-L	Visual Prompt	50.4	54.2	-	-	23.2	15.1	14.3
T-Rex2 [16]	Swin-L	Visual Prompt	-	46.5*	-	45.3*	-	-	-
Prompt-DINO	ViT-L	Text Prompt	51.1	58.6	56.5	60.1	29.6	24.6	28.6
Prompt-DINO	ViT-L	Visual Prompt	51.0	58.0	55.6	59.9	35.9	28.0	32.7

Table 1. Generic Segmentation results of Prompt-DINO on multiple datasets. "-" denotes the model does not have performance reported or does not have ability for the specific task. "*" denotes the performance reported under different domain settings. MaskDINO uses the results reported in DINOv [22]

Datasets. During the training process, we utilized three categories of datasets: *open-source detection datasets.* These include COCO [25] with 118K images and LVIS [13] with 160K images, both of which have mask annotations. For datasets without mask annotations, such as Object365 [38], OID [20], and V3Det [41], we first apply SAM to generate mask annotations. Due to the presence of low-quality data and long-tail distribution in Object365 and OID, which hinder general object recognition, we resampled them, selecting 700K and 1M data points, respectively. *grounding datasets.* This includes the GoldG [17] dataset with 100K images. *self-built datasets.* We re-annotate and filter CC3M [39] and SA-1B [19] using the RAP and dual-path cross-validation data engine introduced in §3.4, yielding high-quality datasets of 3M and 7M images, respectively, which constitute our primary training data. For testing, we utilize COCO and LVIS for in-domain evaluation and ADE20K [55], SeginW [59], and Cityscapes [7] for out-domain evaluation.

Metrics. We employ Average Precision (AP) as the evaluation metric, utilizing mask AP for instance segmentation, and box AP for object detection. For ADE20K and CtiyScapes, we employ Panoptic Quality(PQ) and mean intersection over union (mIoU) for panoptic segmentation as well. For SegInW, we adopt AP-Average and AP-Median as the evaluation metric.

4.2. Generic Segmentation

To evaluate the effectiveness and generalization capability of our Prompt-DINO, we conduct evaluations on both in-domain and out-domain datasets, with the results presented in Tab. 1.

In-Domain Segmentation. We conduct in-domain experiments on COCO and LVIS. Compared to methods based on visual prompts, our Prompt-DINO achieved significantly better results. Compared to the previous work DINOv [22], we improve mask AP by 0.6 and box AP by 3.8 on COCO. Compared to methods based on text prompts, Prompt-DINO also demonstrates superior results when using text prompts. For example, compared to APE [40], we lead by 3.5 in mask AP and 0.5 in box AP on LVIS.

Out-Domain Segmentation. We use ADE-20K and Cityscapes as an out-domain dataset and conduct zero-shot testing on it, demonstrating the strong generalization ability of our Prompt-DINO. Under the out-domain setting (where data marked with * indicates performance reported under different domain settings), our method achieves the best performance on ADE20K across all methods for both prompt types, whether using visual prompts or text prompts. Compared to the state-of-the-art method APE on ADE20K, we improve PQ by 8.7, mask AP by 3.6, and box AP by 18.4, representing increases of 32.0%, 14.8%, and 10.5%, respectively. Among methods based on visual prompts, compared to the previous method DINOv, we improve PQ by 12.7, mask AP by 12.9, and box AP by 18.4, representing increases of 54.7%, 85.4%, and 128.7%, respectively. Compared to APE on Cityscapes, we also improve PQ by 1.0 and mIoU by 8.0.

Segmentation in the Wild. To further evaluate the generalization capabilities of our model, we perform zero-shot testing on the SeginW benchmark, which includes 25 datasets. Compared to the previous work DINOv, when using visual prompts, we improve the average AP by 16.2 and the median AP by 14.5. Among text-prompt-based methods, we also surpass APE by 6.8 in average AP and 6.9 in median AP. This further validates the strong generalization ability of our Prompt-DINO.

4.3. Effectiveness Analysis

To validate the effectiveness of our proposed components, we conduct ablation studies on Text Prompt Alignment, Early Fusion, Order Alignment, and Data Engine. As our baseline, we implement the naive MaskDINO method with the FC head replaced with a user-prompted visual embed-

Method	Type	CityScape		SegInW	
		PQ	mIoU	AP-avg	AP-med
kMaX-DeepLab-L [50]	in-domain	68.4	83.5	-	-
X-Decoder-L [59]	out-domain	38.1	52.0	22.3	32.3
OpenSeeD-L [54]	out-domain	41.4	47.8	36.1	38.7
APE-L [40]	out-domain	33.3	44.2	49.6	52.2
DINOv [22]	out-domain	-	-	40.6	44.6
Prompt-DINO [40]	out-domain	34.3	52.2	56.4	59.1

Table 2. Zero-shot Segmentation results of Prompt-DINO on the CityScape [7] and SegInW [59] dataset.

Method				ADE20K (out-domain)		
Baseline	Text Align	Early Fusion	Order Align	PQ	mAP	bAP
✓	×	×	×	25.0	20.4	23.5
✓	✓	×	×	32.7	25.5	30.5
✓	✓	✓	×	34.7	26.7	31.7
✓	✓	✓	✓	35.9	28.0	32.7

Table 3. **Ablation** of proposed components.

ding. The ablation results are presented in Tab. 3.

Effectiveness of Early Fusion. By comparing rows 1 and 2 in Tab. 3, it can be observed that incorporating text align in the backbone encoder (stage one of DINO) increases the mask AP and box AP on ADE20K by 5.1 and 7.0, respectively, and PQ by 7.7, compared to the baseline. This indicates that naive alignment of text and visual can enhance the model's general recognition capability. Comparing rows 2 and 3 in Tab. 3 shows that early fusion of visual prompts, guided by text alignment, significantly improves the model's performance, with mask AP and box AP on ADE20K increasing by 1.2 and 1.2, respectively, and PQ by 2.0. This demonstrates that visual prompts effectively complement features that are difficult to describe with text prompts and, under the guidance of text prompts, achieve more generalizable expressive power.

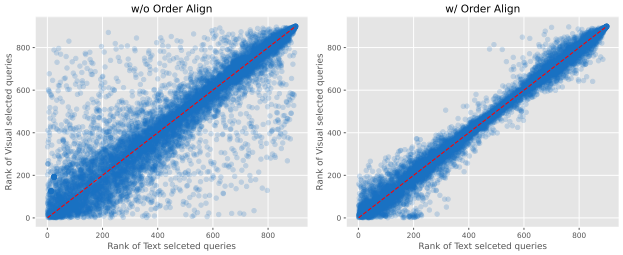


Figure 5. Order Alignment Comparison. **Left:** Visual and text should demonstrate strong correlation, but the queries selected by the two prompts exhibit significant disorder. **Right:** Order alignment significantly improves the consistency of visual-text query selection and reinforces the alignment between corresponding semantics.

Effectiveness of Text-guided Visual Prompts. Rows 1, 2, and 3 in Tab. 3 show that merely adding text alignment provides limited improvement to the model. However, after early fusion to eliminate semantic ambiguity, the two types of prompts can serve as complementary information, providing stronger recognition capabilities. Next, we introduced order alignment during the query selection stage to further guide visual prompts with text. Rows 3 and 4 in Tab. 3 demonstrate that our two-stage guidance deeply

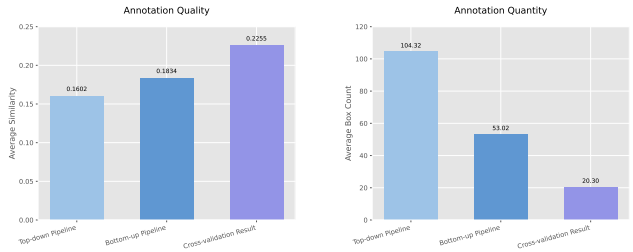


Figure 6. Annotation quality and quantity in the data engine.

aligns visual and text prompts, leading to further performance improvements. Fig. 5 also demonstrates that our order alignment method further enhances the alignment between visual and text, resulting in a more consistent selection order for queries.

Effectiveness of Data Engine. Within the data engine, we employ a cross-validation module to filter data annotated by the top-down and bottom-up pipelines. As illustrated in Fig. 6, the cross-validation module filters approximately 80.5% of instance annotations and enhances the annotation clip similarity from 0.16 to 0.23, highlighting the effectiveness of the cross-validation module in generating high-quality data. As shown in Tab. 4, incorporating data generated by our data engine into the training set improves Prompt-DINO's box AP by 12% on LVIS and 5.9% on ADE20K. To further demonstrate the effectiveness of RAP, we process a subset of SA-1B using different data engines in Tab. 5, revealing that RAP enhances model performance by 1.6% mask AP and 2.6% box AP on ADE20K. This indicates that the proposed data engine effectively enhances data quality, while RAP generates more diverse data, significantly improving the model's generalization capability.

Method	LVIS (in-domain)		ADE20K (out-domain)		
	mask AP	box AP	PQ	mask AP	box AP
w/o DataEngine	46.0	47.9	30.8	23.4	26.8
w Data Engine	55.6	59.9	35.9	28.0	32.7

Table 4. **Ablation** of Data Engine (Visual Prompt Results).

Dataset	COCO (in-domain)		ADE20K (out-domain)		
	mask AP	box AP	PQ	mask AP	box AP
COCO+GoldG	49.8	55.7	18.2	11.0	11.6
COCO+SA-1B (QwenVL)	50.1	56.7	22.3	17.8	20.4
COCO+SA-1B (QwenVL + RAP)	50.8	58.1	22.5	19.4	23.0

Table 5. **Ablation** of Training Dataset (Text Prompt Results). For efficient training, we use a 2M subset of SA-1B.

5. Conclusion

We present Prompt-DINO, a text-guided visual Prompt DINO for generic segmentation via three innovations: 1) Early fusion of text, visual, and backbone features during initial encoding to resolve cross-modal ambiguity; 2) Order-aligned query selection for semantic-spatial consistency between text and visual cues in DETR decoding; 3) A generative data engine (RAP + dual-path verification) synthesizing 0.5B high-fidelity training instances, reducing label noise by 80.5%. Prompt-DINO sets new state-of-the-art results while enabling scalable adaptation to real-world object diversity.

Acknowledgements

This work was partially supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant(KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), National Natural Science Foundation of China (No. 62106149).

References

[1] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 5

[2] 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. 5

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 5

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 7

[5] Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. *arXiv preprint arXiv:2503.08099*, 2025. 2

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 3

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 7, 8

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 4

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 2

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[12] Yuchen Guan, Runxi Cheng, Kang Liu, and Chun Yuan. Enhancing logits distillation with plug&play kendall's τ ranking loss, 2025. 5

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 7

[14] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020. 5

[15] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023. 7

[16] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2024. 2, 3, 4, 7

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 7

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 2

[19] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3, 5, 6, 7

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 7

[21] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 2, 5, 7

[22] Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, et al. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 12861–12871, 2024. 2, 3, 4, 7, 8

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 2, 3, 4, 5

[25] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 7

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2, 3, 4, 5, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[30] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 2

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[32] Ting Pan, Lulu Tang, Xinlong Wang, and Shiguang Shan. Tokenize anything via prompting. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 2

[33] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 4, 6

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 5

[36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[37] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[38] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3, 7

[40] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13193–13203, 2024. 7, 8

[41] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. 7

[42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[43] Yuanrui Wang, Cong Han, YafeiLi, Zhipeng Jin, Xiawei Li, SiNan Du, Wen Tao, Yi Yang, shuanglong li, Chun Yuan, and Liu Lin. Uniglyph: Unified segmentation-conditioned diffusion for precise visual text synthesis, 2025. 2

[44] Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models, 2025.

[45] Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Unifying multimodal large language model capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025.

[46] Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from clip via data-free knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[47] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023. 6

[48] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[49] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pretraining for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2

[50] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 7, 8

[51] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024. 5

[52] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3, 4, 5, 6

[53] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 2, 7

[54] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 7, 8

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 7

[56] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022. 2

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4

[58] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting, 2023. 2

[59] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 7, 8

[60] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 2, 3