TFAG-OVD: TRAINING-FREE ATTRIBUTE-GUIDED OPEN-VOCABULARY DETECTION VIA LLM SEMANTIC COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Currently, open-vocabulary detectors recognize novel concepts primarily by aligning region features with category-name embeddings. However, as the vocabulary expands to thousands of categories, semantically related names crowd the alignment space, making the top match uncertain even at high similarity and disproportionately harming rare or weakly grounded concepts. Therefore, we rethink recognition as attribute-guided object understanding, representing ambiguous regions through transferable, class-agnostic attributes whose compositions can describe objects across datasets and vocabularies. A framework termed Training-Free Attribute-Guided Open-Vocabulary Detection via LLM Semantic Composition (TFAG-OVD) is proposed around a reusable pool of 3,549 class-agnostic attributes. Specifically, a frozen detector produces grounded proposals, and distribution-aware verification preserves clear predictions while routing ambiguous ones for attribute-guided review. For each routed region, a frozen vision–language recognizer retrieves shape, material, part, surface, and state cues; a text-only large language model then composes them within a task-relevant category range to revise the label. This selective attribute interface converts rigid region-to-name matching into evidence-based reasoning while keeping all models frozen. On LVIS minival, TFAG-OVD improves AP from 41.4 to 43.8 and rare-category AP from 34.1 to 37.8 over the frozen detector baseline. Zero-shot results on LVIS val, referring-expression benchmarks, ODinW, COCO, and COCO-O further demonstrate transfer across taxonomies and domains without parameter updates.

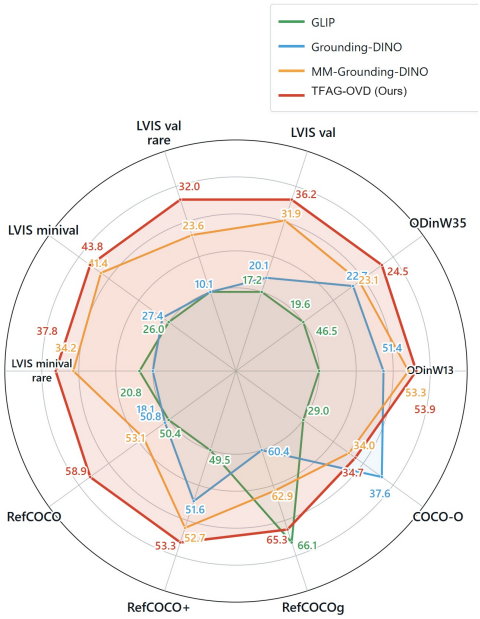


Figure 1: TFAG-OVD achieves superior zero-shot performance across various benchmarks compared with other well-known counterparts. All detectors use Swin-T as the backbone.

1 INTRODUCTION

Current open-vocabulary object detectors recognize novel concepts primarily by aligning region features with category-name embeddings in a shared vision–language space. This direct matching rule scales beyond a fixed training taxonomy, but the alignment space becomes crowded as the vocabulary grows to hundreds or thousands of categories. Semantically adjacent names can therefore receive similar scores, making the highest score unreliable even when its absolute similarity is high. The effect is especially damaging for rare or weakly grounded concepts, where a frequent neighboring class can absorb a visually valid region.

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

The ambiguity is a property of the score distribution, not of one universal cutoff. The complete candidate distribution and the visual properties that separate its plausible interpretations are therefore the relevant evidence. Existing grounded detectors and CLIP-like models provide strong discriminative signals [28, 35, 47], but direct region–name matching does not expose shape, material, parts, surface, or state. We rethink recognition as **attribute-guided object understanding**: transferable, class-agnostic attributes provide an intermediate description whose combinations can characterize objects across datasets and vocabularies. In this view, understanding an arbitrary object amounts to reasoning over and summarizing a composition of its properties.

We present **Training-Free Attribute-Guided Open-Vocabulary Detection via LLM Semantic Composition (TFAG-OVD)**, a framework built around a reusable pool of 3,549 class-agnostic attributes. Specifically, a frozen MM-GroundingDINO detector supplies grounded region candidates through its pretrained zero-shot region–text grounding capability. A distribution-aware ambiguity score evaluates each leading category relative to the full candidate-score distribution, retaining unambiguous predictions and routing uncertain regions for attribute-guided review. This selective routing preserves reliable detector decisions while concentrating additional computation on crowded alignments.

For every routed region, a frozen vision–language attribute recognizer retrieves discriminative cues from the shared pool, including shape, material, parts, surface, and state. A plug-and-play, text-only LLM then composes these cues with a task-relevant detector-derived category range. The language model receives a compact textual description after visual evidence has been extracted; it does not need to attend to redundant low-level pixels, and its output is constrained to plausible detector-supported interpretations. This interface turns rigid region-to-name matching into evidence-based category reasoning while keeping the detector, vision–language model, attribute recognizer, and language model frozen. It also instantiates the open direction identified by Wu et al. [53]: connecting broad LLM knowledge to visual detectors for stronger zero-shot recognition.

On LVIS minival, TFAG-OVD improves AP from 41.4 to 43.8 and rare-category AP from 34.1 to 37.8 over the matched frozen detector baseline. Zero-shot evaluation on LVIS val, referring-expression benchmarks, ODinW, COCO, and COCO-O further tests whether the same attribute interface transfers across taxonomies and domains without parameter updates. Figures 2 and 3 summarize this evidence flow from grounded regions, through uncertainty-aware review and reusable attributes, to language-guided category revision.

Our contributions are:

- **A reusable attribute interface for open vocabularies.** We replace direct region–category matching for ambiguous proposals with a 3,549-entry class-agnostic attribute pool whose compositional descriptions transfer across datasets and taxonomies.
- **Distribution-aware selective reasoning.** We measure whether the leading category is prominent relative to its candidate-score distribution, so modest but isolated matches can be retained while high yet crowded matches receive additional analysis.
- **Semantic composition and broad zero-shot evidence.** A plug-and-play, text-only LLM composes a compact, multi-dimensional attribute description with a task-relevant category range. TFAG-OVD improves Swin-T LVIS minival AP from 41.4 to 43.8 and AP_r from 34.1 to 37.8, and is evaluated on referring expressions, ODinW, COCO, and COCO-O.

2 Related Work

Open-vocabulary detection through alignment. Most OVOD systems replace a closed-set classifier with visual–text similarity. OVR-CNN, ViLD, Detic, OWL-ViT/OWL-ST, RegionCLIP, object-centric alignment, and CORA improve transfer through region-level supervision or pretrained language embeddings [58, 17, 63, 41, 42, 62, 4, 52]. GLIP, GLIPv2, GroundingDINO, and DetCLIP scale this idea with grounding and language-rich pretraining [28, 60, 35, 54, 55, 56], while CapDet, DesCo, YOLO-World, T-Rex2, and OV-DINO broaden the data or prompt space [37, 27, 11, 20, 50]. The common abstraction is still a comparison between a region and a category description. TFAG-OVD targets the ambiguity that appears when many descriptions are simultaneously plausible.

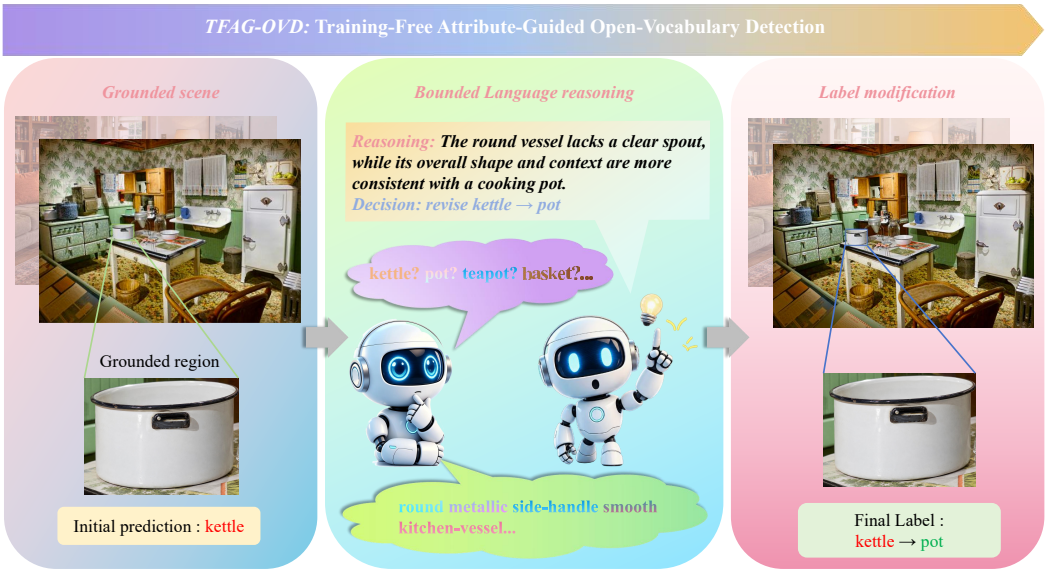


Figure 2: A qualitative view of attribute-guided correction in TFAG-OVD. The frozen detector grounds a region initially labeled *kettle*; a text-only LLM composes the retrieved cues (round, metallic, side handle, and smooth) with a task-relevant set of detector-derived alternatives and revises the label to *pot*. Frozen visual modules remain responsible for localization and attribute evidence.

Attributes as class-agnostic evidence. Attributes expose reusable visual properties that cross category boundaries. OVAD formalizes open-vocabulary attribute recognition, and F-VLM shows that frozen VLM features can classify cropped regions [6, 25]. ComCa makes training-free attribute detection scalable by modeling attribute–object compatibility and assigning compositional soft labels to a cache [16]. The TPAMI survey further identifies parts and attributes as important fine-grained cues when novel classes resemble base classes [53]. We build on this perspective but change the downstream role of attributes: rather than treating them as final predictions, TFAG-OVD uses a substantially expanded class-agnostic pool as a shared descriptive vocabulary. Because its entries express properties rather than dataset labels, their compositions can support different object taxonomies without reconstructing a category-specific attribute table.

Language reasoning for visual recognition. Vision–language assistants such as LLaVA, LLaVA-NeXT, Qwen-VL, Qwen2-VL, CogVLM, InternVL, and Kosmos-2 feed visual tokens or image features to a language model and ask it to reason over the image [33, 34, 2, 3, 51, 10, 45]. Their visual attention is learned for conversational understanding rather than tied to a detector's verified region, so the model may spend capacity on irrelevant context and does not by itself provide a controlled candidate list for each box. LLMDet takes a different route: it distills language-model knowledge into detector training, improving the detector's learned representation but requiring an additional training stage [15]. Text-only foundation models offer a complementary source of broad semantic structure [7, 49, 43, 1, 19, 46]. Wu et al. identify alignment between LLM knowledge and visual detectors as an open direction for stronger zero-shot recognition [53]. TFAG-OVD follows this latter route: frozen visual modules first provide a grounded crop, attributes, and detector-supported alternatives, after which a text-only LLM composes that evidence. The interface separates visual localization from language reasoning, keeps the language model replaceable, and avoids turning it into an unconstrained detector.

Long-tailed recognition and evaluation. LVIS and long-tailed detection motivate separate rare, common, and frequent metrics [18, 31]. Prior work reweights examples, decouples representation and classification, adjusts logits, or calibrates scores after training [12, 22, 40, 44]. TFAG-OVD instead asks whether the semantic evidence for a proposal is sufficiently distinctive before accepting its label. We evaluate this principle with Fixed-AP-style LVIS evaluation [13], referring-expression localization [57, 23, 38, 32], cross-domain ODinW [26], and the COCO-O natural-shift benchmark [30, 39]. These settings complement the detector and backbone foundations inherited from Faster R-CNN, DETR, Deformable DETR, Swin, and ViT [48, 8, 64, 36, 14].

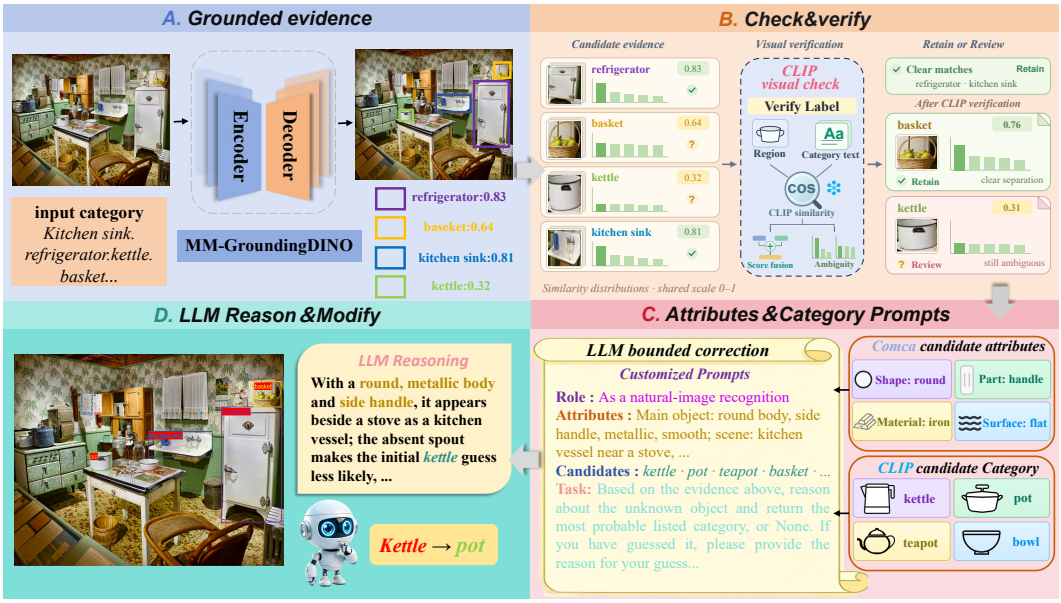


Figure 3: Overview of TFAG-OVD. The frozen MM-GroundingDINO produces grounded regions and candidate evidence. CLIP ViT-L/14 verifies category text and, together with the candidate-score distribution, separates clear matches from ambiguous ones. Ambiguous crops are described by transferable attributes from the shared pool; a frozen text-only LLM composes those cues with a task-relevant category range and revises the label when the evidence supports a different interpretation.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Given an image I and an inference-time vocabulary $\mathcal{C}$, TFAG-OVD predicts region boxes, confidence scores, and categories. Direct alignment chooses a category from region–name similarities but does not explain which observable properties distinguish it from nearby categories, so ambiguous or long-tail objects may receive plausible but incorrect names. TFAG-OVD introduces attribute evidence $\mathcal{A}_i$ only for ambiguous decisions: distinctive predictions follow the direct path, while crowded ones follow the attribute-guided path before reconsideration. Every output remains in $\mathcal{C}$ and every model is frozen.

Figure 3 maps the information flow to panels A–D. Panel A grounds region proposals with frozen MM-GroundingDINO; panel B retains clear category distributions and routes ambiguous ones for review. Panel C pairs attributes from each routed crop with detector-supported alternatives, and panel D lets the frozen text-only LLM compose this evidence and revise the label when warranted. Direct and revised predictions are finally merged by common score calibration and overlap suppression. Thus, language reasoning follows visual evidence extraction rather than replacing localization or expanding the inference vocabulary.

Figure 3 illustrates why an absolute similarity is insufficient: a leading score of 0.70 can be decisive when the remaining candidates are far lower, whereas a score of 0.99 can remain ambiguous when several candidates are nearly tied.

3.2 GROUNDED REGION EVIDENCE

TFAG-OVD adopts frozen MM-GroundingDINO [61] because large-scale multimodal pretraining equips it with zero-shot region–text grounding and proposal inference. Given image I and category prompts $\mathcal{C}$, detector D_ϕ produces

$$\mathcal{P}^0 = D_\phi(I, \mathcal{C}) = \{(b_i, s_i, c_i)\}_{i=1}^{N}, \quad c_i = \arg\max_{c \in \mathcal{C}} d_{i,c}, \quad s_i = d_{i,c_i}, \tag{1}$$

where ϕ is frozen, $d_{i,c}$ is the grounding score between region query i and category c, b_i is its box, and $x_i = \mathrm{Crop}(I, b_i)$ is its downstream crop. Text-budgeted query groups are merged before class-wise consolidation removes duplicate instances. Cross-class alternatives at one location are retained

because their joint score profile reveals whether an interpretation is distinctive or crowded. No detector parameter is updated on the evaluation taxonomy.

3.3 Distribution-Aware Verification

We use frozen CLIP ViT-L/14 features [47] as an independent verifier rather than immediately replacing the detector label. For each category c, its name and lexical definition form the prompt

$$t_c = \text{“a photo of a } \mathrm{name}(c), \mathrm{def}(c)\text{”}. \tag{2}$$

Here $\mathrm{name}(c)$ is the category name and $\mathrm{def}(c)$ is its lexical definition from metadata; it enriches the category prompt but is not an attribute prediction. Let $v_{i,c}$ be the cosine similarity between proposal crop x_i defined above and prompt t_c, and let $m_i = v_{i,c_i}$ denote agreement with the detector label. Detector and verifier evidence are fused as

$$s_i^* = (1-\alpha)s_i + \alpha \ \mathrm{clip}\left(\frac{m_i - \ell}{h - \ell}, 0, 1\right). \tag{3}$$

Here s_i is the detector score, m_i is the CLIP ViT-L/14 cosine similarity for the detector label, α is the fusion weight, and ℓ and h are the lower and upper calibration bounds. The operator $\mathrm{clip}(z, 0, 1) = \min(1, \max(0, z))$ limits the normalized verifier score, and s_i^* is the resulting fused proposal score. For routing, the comparable evidence is the verifier row itself: let $v_{i,(1)} \geq \cdots \geq v_{i,(k)}$ be the top-k category similarities, let $\Delta_i = v_{i,(1)} - v_{i,(2)}$ be the leading margin, and compute

$$\mu_i = \frac{1}{k}\sum_{j=1}^{k} v_{i,(j)}, \qquad \sigma_i = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(v_{i,(j)} - \mu_i)^2}. \tag{4}$$

We then define normalized evidence terms $a_i = \mathrm{clip}((v_{i,(1)} - \ell)/(h - \ell), 0, 1)$, $d_i = \mathrm{clip}(\Delta_i/(h - \ell), 0, 1)$, and $z_i = \mathrm{clip}((v_{i,(1)} - \mu_i)/(\sigma_i + \varepsilon), 0, 1)$, where ε prevents division by zero. We compute these statistics from the verifier row rather than s_i^*: the latter is one fused score for the detector label and cannot describe competition among candidate categories. The distribution-aware ambiguity score is

$$u_i = 1 - \frac{w_a a_i + w_d d_i + w_z z_i}{w_a + w_d + w_z}, \qquad w_a, w_d, w_z > 0. \tag{5}$$

We use equal weights unless stated otherwise. Similar matches often yield a small leading margin and may produce a compact top-k distribution, but variance alone is insufficient because uniformly low and uniformly high rows can have similar spread. Hence u_i is low only when the leader is strong, separated, and atypical within its row, and high when evidence is weak or crowded. Proposals with $u_i \geq \tau_u$ enter the attribute-guided path; the remainder retain the verified label. The verifier distribution governs routing, while s_i^* remains the ranking and fusion score.

3.4 Class-Agnostic Attributes and Category Prompts

The main representation in TFAG-OVD is a shared pool $\mathcal{A}$ of 3,549 class-agnostic attributes. We collect lexical concepts from VAW and PACO attribute annotations, WordNet descriptors, and ComCa-derived concepts; normalize spelling and formatting, merge duplicate or synonymous entries, and retain terms that denote observable evidence. The resulting pool is organized into five complementary families: appearance and surface, geometry and topology, material and physical properties, parts and components, and state or context. These terms describe evidence without assuming a category label, allowing the same cue—for example, `round`, `metallic`, or `has_handle`—to support many objects. Their combinations remain meaningful when the inference taxonomy or dataset changes, so the same pool is reused throughout our zero-shot evaluations; construction and semantic organization details are provided in Appendix B.

For an ambiguous crop x_i, a frozen OpenCLIP attribute recognizer scores every attribute prompt,

$$r_{i,a} = \cos(f_v(x_i), f_t(\text{“a photo of a } a \text{ object”})), \qquad a \in \mathcal{A}. \tag{6}$$

Following ComCa's compositional view of attributes [16], the highest-scoring compatible cues form $\mathcal{A}_i$. The category similarities used for the candidate range are a separate output of the frozen MM-GroundingDINO/CLIP category-verification path: they are scored against the inference-time vocabulary $\mathcal{C}$ and are not mapped from the attribute scores. The top-k entries form a task-relevant category

range $\mathcal{K}_i \subset \mathcal{C}$, with the detector label included if necessary. The resulting prompt states what is visible and which taxonomic interpretations remain plausible, allowing the language model to compose evidence rather than inspect pixels or invent an unrelated category.

3.5 Language-Guided Category Revision

A frozen text-only LLM receives $(\mathcal{A}_i, \mathcal{K}_i)$ and returns

$$\hat{c}_i = g_\theta(\mathcal{A}_i, \mathcal{K}_i) \in \mathcal{K}_i \cup \{\perp\}, \tag{7}$$

where θ denotes the frozen LLM parameters, $\mathcal{A}_i$ is the selected attribute evidence, $\mathcal{K}_i$ is the detector-derived task-relevant category range, $\hat{c}_i$ is the proposed revised category, and $\perp$ denotes abstention. The prompt and output parser accept only a member of $\mathcal{K}_i$ or abstention. A valid alternative revises the ambiguous label; agreement or abstention preserves the detector decision. This textual interface keeps low-level visual detail out of the language decision and prevents unconstrained category hallucination. Its contribution is semantic composition—relating shape, material, parts, and context to a name—rather than visual localization.

The direct and attribute-guided paths are finally merged. Label revision preserves the fused proposal score, so let $q_i = s_i^*$ denote the score entering the final calibration. A light rare-category calibration inspired by NORCAL [44] adjusts long-tail scores,

$$q_i \leftarrow \min(1, q_i\, \lambda_r^\gamma), \tag{8}$$

and Gaussian Soft-NMS [5] consolidates overlapping detections,

$$q_j \leftarrow q_j \exp\left(-\frac{\mathrm{IoU}(b_j, b_*)^2}{\sigma}\right). \tag{9}$$

Here λ_r is the rare-category calibration factor, γ is its exponent, q_j is the score of an overlapping box, b_* is the currently selected highest-score box, IoU is intersection-over-union, and σ controls the Gaussian decay. The essential change is the reusable attribute interface: an uncertain visual match is converted into class-agnostic evidence before a category is reconsidered.

4 Experiments

4.1 Implementation Details

We evaluate LVIS v1.0 minival (4,809 images) and val (19,809 images) [18]. The evaluator follows the Fixed-AP code path [13] with `max_dets=-1`. The complete Swin-T system uses the public MM-GroundingDINO Swin-T checkpoint in MMDetection [9, 61], CLIP ViT-L/14 for category verification, OpenCLIP ViT-B/32 for attribute recognition, and Qwen3.5-9B as the locally served text-only classifier [46]. The reported operating point uses fusion weight $\alpha = 0.35$, verifier calibration bounds $(\ell, h) = (0.10, 0.35)$, ambiguity threshold $\tau_u = 0.30$, equal routing weights, 15 returned attributes, and $k = 15$ category candidates. Rare-category calibration uses $(\lambda_r, \gamma) = (1.5, 0.3)$, and Soft-NMS uses $\sigma = 0.5$. Qwen3.5-9B is an experimental backend for reproducibility rather than part of the method definition; another capable frozen text-only LLM can instantiate the same category-revision interface. All evaluations are zero-shot: no detector, VLM, attribute model, or LLM parameters are updated on evaluation data. We additionally evaluate transfer on ODinW, COCO, COCO-O, and referring-expression benchmarks. Unless stated otherwise, benchmark metrics are single-run percentages rounded to one decimal; literature rows retain their published protocol.

4.2 Comparison Experiments

Following the broad-coverage evaluation practice used in recent OVOD work [15], we report zero-shot results on LVIS, referring-expression localization, ODinW, and COCO/COCO-O. Together, these benchmarks test long-tailed recognition, language-conditioned localization, cross-domain transfer, and natural distribution shift while reusing the same class-agnostic attribute pool. Literature rows follow their published protocols, whereas TFAG-OVD rows use the frozen checkpoints and benchmark-specific evaluation described above. Each table is therefore interpreted within its stated protocol rather than as a matched rerun of every prior method.

LVIS. Table 1 reports AP and the frequency-stratified LVIS metrics for 19 distinct prior methods spanning convolutional, ViT, and Swin backbones. The training-data column identifies methods that use LVIS base-class boxes, while every literature row retains its published evaluation protocol. Under the matched public MM-GroundingDINO Swin-T checkpoint and the same evaluator, **TFAG-OVD improves minival AP from 41.4 to 43.8 and AP_r from 34.1 to 37.8**; on val it reaches 36.2 AP and 32.0 AP_r. The larger rare-category improvement is consistent with the proposed motivation: descriptions based on parts, materials, shape, surface, and state remain informative when a category name has weak direct support.

Table 1: LVIS open-vocabulary comparison. Mini/Val report box AP/AP_r/AP_c/AP_f (%) at one-decimal precision. Lavender marks TFAG-OVD and bold marks the largest displayed value. $LVIS_b$ denotes base-class box supervision without rare-class boxes; G denotes GoldG; dashes are unreported.

			LVIS minival				LVIS val			
Method	Backbone	Training data	AP	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f
MDETR [21]	RN101	GoldG+	24.2	20.9	24.9	24.3	–	–	–	–
ViLD-Ens. [17]	EN-B7	ALIGN,$LVIS_b$	–	–	–	–	31.8	27.0	29.4	36.5
RegionCLIP [62]	R50x4	CC3M,$LVIS_b$	–	–	–	–	32.3	22.0	32.1	36.9
OWL-ViT [41, 42]	CLIP-L/14	O365,VG	–	–	–	–	34.6	31.2	–	–
Detic [63]	R50	IN-L,$LVIS_b$	–	–	–	–	33.4	26.7	–	–
GLIP [28]	Swin-T	O365,G,Cap4M	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.2
GLIPv2 [60]	Swin-T	O365,G,Cap4M	29.0	–	–	–	–	–	–	–
CapDet [37]	Swin-T	O365,VG	33.8	29.6	32.8	35.5	–	–	–	–
Grounding-DINO [35]	Swin-T	O365,G,Cap4M	27.4	18.1	23.3	32.7	20.1	10.1	15.3	29.9
RO-ViT [24]	ViT-B/16	ALIGN,$LVIS_b$	–	–	–	–	31.9	28.4	–	–
OWL-ST [42]	CLIP-B/16	WebLI-2B	34.4	**38.3**	–	–	28.6	30.3	–	–
DesCo-GLIP [27]	Swin-T	O365,G,CC3M	34.6	30.8	30.5	39.0	26.2	19.6	22.0	33.6
DetCLIP [54]	Swin-T	O365,G,YFCC1M	35.9	33.2	35.7	36.4	28.4	25.0	27.0	28.4
DetCLIPv2 [55]	Swin-T	O365,G,CC15M	40.4	36.0	41.7	40.4	32.8	31.0	31.7	34.8
GenerateU [29]	Swin-T	VG,GRIT5M	26.8	20.0	24.9	29.8	–	–	–	–
YOLO-World-L [11]	YOLOv8-L	O365,G,CC3M	35.4	27.6	34.1	38.0	–	–	–	–
T-Rex2 [20]	Swin-T	10M mixed	42.8	37.4	39.7	**46.5**	34.8	29.0	31.5	41.2
OV-DINO [50]	Swin-T	O365,G,CC1M	40.1	34.5	39.5	41.5	32.9	29.1	30.4	37.4
MM-GDINO [61]	Swin-T	O365,G,GRIT,V3Det	41.4	34.1	37.4	46.3	31.9	23.6	27.6	40.5
TFAG-OVD (ours)	Swin-T	O365,G,GRIT,V3Det	**43.8**	37.8	**42.1**	**46.5**	**36.2**	**32.0**	**32.3**	**42.5**

Referring expressions. Table 2 evaluates zero-shot referring-expression localization on RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO. TFAG-OVD improves over MM-GDINO on all three RefCOCO splits, including a **+5.8 Top-1 gain on RefCOCO val**. These expression-defined queries differ from LVIS category names, so the improvement provides evidence that the shared attributes function as reusable visual descriptions rather than LVIS-specific labels.

Table 2: Zero-shot referring-expression transfer. RefCOCO, RefCOCO+, and RefCOCOg report Top-1 accuracy (%). gRefCOCO reports Pr(F1=1, IoU$\geq$0.5)/N-acc. Lavender marks TFAG-OVD; bold marks each column maximum.

		RefCOCO [57]			RefCOCO+ [57]			RefCOCOg [38]		gRefCOCO [32]		
Method	Backbone	val	testA	testB	val	testA	testB	val	test	val	testA	testB
GLIP (B) [28]	Swin-T	50.0	54.7	43.1	49.0	53.4	43.4	65.6	66.1	–	–	–
GLIP [28]	Swin-T	50.4	54.3	43.8	49.5	52.8	44.6	**66.1**	**66.9**	–	–	–
Grounding-DINO [35]	Swin-T	50.8	57.4	45.0	51.6	57.3	46.4	60.4	59.7	40.5/83.8	**29.3**/82.9	30.0/86.1
MM-GDINO [61]	Swin-T	53.1	59.1	46.8	52.7	58.7	48.4	62.9	62.9	41.0/91.3	26.1/93.0	30.4/92.3
TFAG-OVD (ours)	Swin-T	**58.9**	**64.5**	**53.1**	**53.3**	**58.8**	**48.9**	65.3	64.4	**59.5/91.9**	26.8/**93.2**	**31.4/92.8**

ODinW. On the zero-shot ODinW transfer benchmark, TFAG-OVD obtains the best displayed values for both the 13-dataset and 35-dataset aggregates (Table 3), improving over MM-GDINO by 0.6 and 1.4 AP, respectively. These heterogeneous domains test the same reusable attribute pool beyond the label distribution that motivated it.

Table 3: Zero-shot transfer on ODinW [26]. ODinW13 and ODinW35 average 13 and 35 datasets, respectively; values are AP (%).

Method	Backbone	ODinW13	ODinW35
MDETR [21]	EN-B5	–	10.7
T-Rex2 [20]	Swin-T	–	18.0
OWL-ViT [42]	ViT-L/14	53.0	18.8
GLIP [28]	Swin-T	46.5	19.6
GLIPv2 [60]	Swin-T	48.5	22.3
DetCLIP [54]	Swin-T	43.3	–
Grounding-DINO [35]	Swin-T	51.4	22.7
MM-GDINO [61]	Swin-T	53.3	23.1
TFAG-OVD (ours)	Swin-T	**53.9**	**24.5**

COCO and COCO-O. Table 4 completes the zero-shot evaluation on standard COCO and the natural-shift COCO-O benchmark. TFAG-OVD preserves the strong COCO baseline and improves COCO-O AP from 34.0 to 34.7. This result examines the shared attribute interface under a standard taxonomy and natural distribution shifts rather than only a long-tailed label space.

Table 4: Zero-shot COCO and COCO-O [30, 39] comparison. Values are AP (%); COCO-O covers natural distribution shifts.

Method	Backbone	COCO	COCO-O
DINO [59]	R50	49.0	22.5
GLIP [28]	Swin-T	46.1	29.0
DetCLIPv3 [56]	Swin-T	47.2	**38.5**
Grounding-DINO [35]	Swin-T	48.4	37.6
MM-GDINO [61]	Swin-T	50.4	34.0
TFAG-OVD (ours)	Swin-T	**50.5**	34.7

4.3 Ablation Study

4.3.1 Main Components and Rare Categories

All ablations are zero-shot inference experiments on LVIS minival. Table 5 jointly reports AP and rare-class AP_r while isolating proposal consolidation, category verification, and the coupled attribute–language branch; the method names describe which components are enabled on top of the MM-GroundingDINO detector. Attributes and the LLM are evaluated together because the attributes are the evidence consumed by the language classifier. NMS alone slightly lowers AP and AP_r (41.4/34.0 to 41.1/33.4), showing that consolidation can remove useful cross-class alternatives when it is used without semantic verification. CLIP verification alone raises AP to 42.6 and AP_r to 34.7, while the attribute–language branch alone reaches 42.9/35.1, indicating complementary gains from visual category checking and transferable property descriptions. Combining NMS and CLIP yields 43.3/36.6; adding the coupled branch to this variant gives the full 43.8/37.8, a further **+0.5 AP and +1.2 AP_r**. When CLIP verification is disabled, the attribute–language variants route all proposals to that branch, whereas the full path first uses the score distribution to select ambiguous regions; these settings therefore measure different routing conditions. The complete system improves the detector baseline by **+2.4 AP and +3.8 AP_r**, with the larger rare-category gain consistent with class-agnostic attributes supplying evidence when direct category-name matching is weak.

Table 5: Component ablation and rare-category improvement on LVIS minival. Checkmarks denote active components; AP and AP_r (%) are rounded to one decimal.

MM-GDINO	Proposal NMS	CLIP verify	Attr.+LLM	AP	AP_r
✓				41.4	34.0
✓	✓			41.1	33.4
✓		✓		42.6	34.7
✓			✓	42.9	35.1
✓	✓	✓		43.3	36.6
✓	✓		✓	43.5	37.0
✓		✓	✓	43.7	37.5
✓	✓	✓	✓	**43.8**	**37.8**

4.3.2 HYPERPARAMETER SENSITIVITY

We vary one parameter at a time on LVIS minival while holding the remaining settings fixed. These runs test the operating point rather than statistical significance.

CLIP fusion coefficient. Table 6 shows a non-monotonic dependence on α. The selected value, 0.35, gives the highest measured AP; the sharp drop at 1.0 establishes that CLIP crop similarity cannot replace the detector score.

Table 6: CLIP fusion-weight sensitivity on LVIS minival. The remaining settings are fixed; AP values are percentages.

α	0.00	0.20	**0.35**	0.50	1.00
AP	41.1	43.4	**43.8**	42.7	29.9

Ambiguity-routing threshold. We conduct this experiment because the routing threshold determines how much evidence is delegated to attribute-guided review. A threshold that is too low reviews many already separable regions, whereas a threshold that is too high leaves crowded distributions on the direct path. Table 7 varies τ_u on the distribution-aware ambiguity score; lower values broaden review and higher values reserve it for more ambiguous distributions. Each entry is obtained by evaluating a complete pipeline run at that threshold. Both AP and AP_r rise from $\tau_u = 0.10$ to a joint maximum at 0.30 (43.8 and 37.8), then decline as the threshold increases; the measured spans are 2.5 and 2.5 points, respectively. We therefore use $\tau_u = 0.30$ for the updated operating point.

Table 7: Distribution-aware ambiguity-threshold sensitivity on LVIS minival. All benchmark metrics are percentages rounded to one decimal; bold marks the selected operating point and its corresponding results.

τ_u	AP (%)	AP_r (%)
0.10	42.2	36.1
0.20	43.3	36.9
0.30	**43.8**	**37.8**
0.40	43.5	37.3
0.50	42.8	36.5
0.60	41.7	35.9
0.70	41.3	35.3

Candidate-range sensitivity appears in Appendix C.

5 CONCLUSION

TFAG-OVD replaces uncertain region–name matching with transferable attribute evidence. Distribution-aware routing selects ambiguous predictions, and a frozen text-only LLM composes retrieved attributes within a task-relevant category range. Under matched LVIS evaluation, **Swin-T minival AP rises from 41.4 to 43.8 and AP_r from 34.1 to 37.8**; zero-shot transfer holds across taxonomies and domains.

AI ASSISTANCE DISCLOSURE

The authors used AI tools only to aid or polish the writing by assisting with translation and improving the manuscript's grammar, word choice, and readability; all AI-assisted text was subsequently reviewed and revised by the authors.

REFERENCES

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Junyang Bai, Shuai Li, Shusheng Zhang, Jinze Chen, Peng Bai, Yunfei Chu, Xiaodong Wang, Chang Zhou, and Jingren Zhou. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[4] Monika Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33781–33794, 2022.

[5] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS – improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5561–5569, 2017.

[6] María A Bravo, Sudhanshu Mittal, Thomas Brox, and Jürgen Gall. OVAD: Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7338–7348, 2023.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[11] 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*, 2024.

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

[13] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021.

[14] 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. In *International Conference on Learning Representations*, 2021.

[15] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*, 2025.

[16] Marco Garosi, Alessandro Conti, Gaowen Liu, Elisa Ricci, and Massimiliano Mancini. ComCa: Compositional caching for training-free open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15098–15107, 2025.

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

[18] 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*, pp. 5356–5364, 2019.

[19] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[20] 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*, 2024.

[21] 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*, pp. 1780–1790, 2021.

[22] Bo Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Zheng Shou. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 787–798, 2014.

[24] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11144–11154, 2023. doi: 10.1109/CVPR52729.2023.01072.

[25] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023.

[26] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *Advances in Neural Information Processing Systems*, 2022.

[27] Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. DesCo: Learning object recognition with rich language descriptions. In *Advances in Neural Information Processing Systems*, 2023.

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. GLIP: Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

[29] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13958–13968, 2024. doi: 10.1109/CVPR52733.2024.01324.

[30] 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 *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[32] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge. *arXiv preprint arXiv:2401.13601*, 2024.

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Junfeng Gao, Jianfeng Wang, Wenlong Xiang, Chunyuan Li, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

[36] 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*, pp. 10012–10022, 2021.

[37] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. CapDet: Unifying dense captioning and open-world detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15233–15243, 2023.

[38] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

[39] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. COCO-O: A benchmark for object detectors under natural distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[40] Aditya Krishna Menon, Sadeepa Jayasumana, Anil Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[41] 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*, pp. 728–755. Springer, 2022.

[42] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

[43] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[44] Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On model calibration for long-tailed object detection and instance segmentation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[45] 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. In *Advances in Neural Information Processing Systems*, 2023.

[46] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[47] 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*, pp. 8748–8763. PMLR, 2021.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[50] Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. OV-DINO: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024.

[51] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[52] Jiahao Wu, Qiang Shen, Wenhao Huang, Zhongtian Guo, Ziyun Li, Yifan Lu, Jifeng Dai, Yu Qiao, Hongsheng Li, and Xiaodan Li. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7031–7040, 2023.

[53] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5092–5113, 2024. doi: 10.1109/TPAMI.2024.3361862.

[54] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Hu. DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9125–9138, 2022.

[55] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Chunjing Xu. DetCLIPv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23497–23506, 2023.

[56] Lewei Yao, Yuntao Huang, Jiahao Zhang, Jianhua Han, Chunjing Xu, Huifeng Wang, Wei Zhang, Zhenguo Li, and Dan Xu. DetCLIPv3: Towards versatile generative open-vocabulary object detection. *arXiv preprint arXiv:2404.09216*, 2024.

[57] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

[58] Alireza Zareian, Kevin Duarte Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

[59] 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. In *International Conference on Learning Representations*, 2023.

[60] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36067–36080, 2022.

[61] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024.

[62] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

[63] 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*, pp. 350–368. Springer, 2022.

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

APPENDIX

A COMPLETE HYPERPARAMETERS

Table 8: Configuration of the reported TFAG-OVD minival run. All listed models are frozen.

Component	Parameter	Value
Grounded detector	score threshold / chunk size / max per chunk	0.05 / 40 / 300
Proposal consolidation	IoU threshold / max per class	0.60 / 50
Visual verification	CLIP model / views / fusion α	ViT-L/14 / 3 / 0.35
Attribute evidence	ambiguity threshold / weights / attributes / cutoff	0.30 / equal / 15 / 0.15
Language classification	candidates / model / score / tokens	15 / Qwen3.5-9B / 0.15 / 200
Detection merge	rare γ / Soft-NMS σ / cutoff	0.30 / 0.50 / 0.03

B ATTRIBUTE POOL AND COMCA INFERENCE

The reusable attribute pool contains 3,549 normalized entries collected from VAW, PACO, WordNet, and ComCa-derived concepts. We lower-case and normalize spelling, replace separators with spaces, merge duplicates and near-synonyms, and retain descriptors that refer to observable evidence. The final inventory is organized into five families: appearance/surface, geometry/topology, material/physical properties, parts/components, and state/context. This organization is semantic rather than category-specific, so a descriptor such as "round", "metallic", or "has handle" can participate in descriptions for multiple object categories. Table 9 summarizes the construction process and illustrates how representative descriptors provide reusable evidence across categories.

For each attribute a, the frozen OpenCLIP attribute recognizer encodes the prompt "a photo of a a object" once and caches the normalized text embedding. A pending detector box is clipped to the image boundary and cropped without adding context padding in the reported LVIS run. Its RGB crop is encoded into a normalized 512-dimensional visual feature, and the dot products with all 3,549 attribute embeddings provide the visual attribute scores. We sort these scores, keep at most 15 attributes above 0.15, and retain the highest-scoring attribute if all scores fall below the cutoff. The resulting list is the textual evidence supplied to the language classifier.

ComCa-style support caches are optional in this path. When enabled, support-image keys and soft attribute values add a cache logit to the visual–attribute scores; without a cache the scorer reduces to the OpenCLIP cosine similarities. A separately generated 3549×1203 compatibility matrix expresses language-level attribute–category relatedness. It is a prior for variants that explicitly pass it to the inferencer and is not conflated with the visual crop scores. The reported main LVIS path uses the shared pool and visual attribute scores, with cache usage controlled by the run configuration.

The complete data flow is therefore: image and inference vocabulary $\rightarrow$ frozen MM-GroundingDINO proposals $\rightarrow$ CLIP category-score rows and distribution-aware routing $\rightarrow$ XYXY crop for routed boxes $\rightarrow$ OpenCLIP attribute scores $\rightarrow$ top-15 attribute evidence $\rightarrow$ task-relevant category range $\rightarrow$ frozen text-only LLM revision or abstention. The LLM receives only the selected attribute strings and detector-derived category names; agreement or abstention leaves the detector label unchanged.

Table 9: Construction and semantic organization of the shared class-agnostic attribute pool. Panel (a) summarizes the curation pipeline; panel (b) illustrates the cross-category reuse of representative visual descriptors.

(a) Attribute-pool construction

Stage	**Source or operation**	**Result**
Collection	VAW, PACO, WordNet, and ComCa-derived concepts	Candidate descriptors
Normalization	Lower-case; normalize spelling; replace separators with spaces	Standardized descriptors
Consolidation	Merge duplicate entries and near-synonyms	Reduced lexical redundancy
Filtering	Retain descriptors that refer to observable evidence	Observable attribute descriptors
Organization	Group descriptors into five semantic families	3,549 class-agnostic attributes

(b) Semantic families and cross-category reuse

Attribute family	**Representative descriptor**	**Visual evidence**	**Illustrative categories**
Appearance / surface	smooth	Surface appearance and texture	pot, bowl, bottle
Geometry / topology	round	Shape and geometric structure	pot, plate, ball
Material / physical properties	metallic	Material and physical properties	kettle, pot, spoon
Parts / components	has handle	Observable parts and components	kettle, pot, basket
State / context	–	Observable states or contextual cues	–

C CANDIDATE-RANGE SENSITIVITY

The category range must contain plausible semantic alternatives without overwhelming the language prompt with unrelated names. The sweep below reports both AP and AP_r; values are percentages rounded to one decimal. Bold values denote the selected operating point ($k = 15$) used in all main experiments.

Table 10: Task-relevant category-range sensitivity on LVIS minival.

Top-k candidates	3	5	10	**15**	20
AP (%)	41.4	42.5	43.2	**43.6**	43.8
AP_r (%)	35.3	36.4	37.1	**37.6**	38.1

D REPRODUCIBILITY NOTES

The complete minival block is AP=43.8, AP_{50}=55.1, AP_{75}=46.1, AP_r=37.8, AP_c=42.1, and AP_f=46.5. The matched baseline block is AP=41.4, AP_{50}=51.8, AP_{75}=43.9, AP_r=34.1, AP_c=37.4, and AP_f=46.3. Tables 1 and 5 report LVIS comparison results at one-decimal precision, and this reproducibility note follows the same rounding convention. The complete-run block and the sensitivity sweeps are separate recorded evaluations; the reported main configuration retains $\alpha = 0.35$, $\tau_u = 0.30$, and top-15 category candidates.

Limitations and future directions. The current system incurs attribute-recognition and LLM inference cost, uses a manually assembled pool that may omit domain-specific or non-visual concepts, and evaluates the attribute recognizer and LLM as a coupled branch. Cross-paper comparisons should also be interpreted with care because the methods reported in prior work use different pre-training protocols. Automatic pool expansion, structured synonym handling, task-adaptive retrieval, and extension to video, 3D scenes, or continually changing vocabularies remain future directions.