A Survey on Knowledge Distillation of Vision-Language Models

Anonymous ACL submission

Abstract

Vision-language models (VLMs) achieve strong multimodal performance but remain expensive to deploy. Knowledge distillation (KD) is an effective approach for compressing these models and transferring their capabilities to more efficient student models. While KD for large language models (LLMs) is already well surveyed, no existing work provides systematic coverage of KD for VLMs, where the addition of a visual modality introduces challenges that text-only distillation does not face. This paper presents the first dedicated survey of knowledge distillation for vision-language models. We propose a unified taxonomy that organizes methods along three orthogonal views: knowledge type, teacher accessibility, and structural compatibility, use it to organize methods across major VLM distillation settings. We also survey the associated datasets and evaluation protocols, identify open challenges and outline directions for future research.

1 Introduction

Vision-language models (VLMs) integrate visual encoders with language model backbones to enable tasks requiring joint visual-linguistic understanding. Recent models such as LLaVA (Liu et al., 2023a), InternVL (Chen et al., 2024d), Qwen-VL (Bai et al., 2025), and DeepSeek-VL (Wu et al., 2024c) achieve strong results on multimodal benchmarks, but their computational cost, combining billion-parameter visual encoders with large language backbones, limits deployment on resource-constrained devices. Knowledge distillation (KD) (Hinton et al., 2015) addresses these limits by training a compact student model to approximate a stronger teacher's behavior, retaining much of its capability at a substantially smaller size. While KD for text-only large language models (LLMs) is now well surveyed (Xu et al., 2024b; Yang et al., 2025a; Moslemi et al., 2024; Fang et al., 2026), no existing survey provides systematic coverage of KD for VLMs, where the addition of a visual modality introduces challenges that text-only distillation does not face. KD for VLMs has consequently grown into an active research area, driven by demands across several fronts. The direct objective is compressing large multimodal models into compact students that retain multimodal reasoning capabilities at a fraction of the original cost while preserving conversational quality and visual grounding (Wang et al., 2023b; Shu et al., 2024; Cai et al., 2025; Feng et al., 2025b; Lee et al., 2025a). Beyond compression, KD transfers VLM representations into task-specific architectures: open-vocabulary detection (Wang et al., 2023a; Li et al., 2023b), zero-shot segmentation (Xiao et al., 2024), embodied navigation (Sumers et al., 2023a), and vision-based reinforcement learning (Xu et al., 2025). A parallel line of work compresses contrastive VLMs for mobile deployment through affinity mimicking (Wu et al., 2023b), cluster-instance discrimination (Yang et al., 2025b), or embedding-level transfer (Nair, 2024; Yang et al., 2024). Furthermore, self-distillation operates during VLM pretraining to strengthen cross-modal alignment without an external teacher (Dong et al., 2023; Jing et al., 2024; Kim et al., 2025b). The broad range of these settings, each with distinct objectives, knowledge signals, and evaluation criteria, motivates a systematic survey.

Despite this growth, no existing survey provides dedicated and in-depth coverage of KD for VLMs. Existing surveys treat distillation only peripherally: (Zhang et al., 2024a) devote one section to distilling VLM knowledge into downstream tasks, without covering generative VLM compression, CLIP distillation, or cross-modal feature alignment methods. (Shinde et al., 2025) survey efficient VLM techniques and include KD as one subsection, but do not provide a taxonomy or systematic method-level analysis of distillation approaches. Broader

multimodal surveys (Yin et al., 2024; Liang et al., 2024) cover VLM architectures, training, and evaluation, not treat distillation as a focus area. On the other hand, KD surveys do not address the multimodal setting: LLM-specific surveys (Xu et al., 2024b; Yang et al., 2025a; Fang et al., 2026) cover text-only distillation, and general KD surveys (Gou et al., 2021; Mansourian et al., 2025) mention VLMs only briefly. The result is that a researcher entering VLM distillation still lacks a single reference that systematically organizes the field: which methods exist, how they relate to each other, what knowledge signals they transfer, what teacher-access assumptions they require, how structural mismatches are handled, which evaluation settings are comparable across applications, and what problems remain open. This paper addresses this gap by providing the **first dedicated survey of knowledge distillation for vision-language models**. We observe that every VLM distillation method must answer three independent questions: **(1)** what knowledge is transferred from teacher to student, **(2)** what level of access to the teacher is available, and **(3)** what structural mismatch must be bridged before transfer can occur. These questions define a taxonomy that enables systematic cross-method comparison. To ensure transparency, **our literature collection methodology is detailed explicitly in Appendix B**. We also survey **datasets, evaluation protocols and their evolution in Appendix D** across taxonomy categories. Our main contributions are: **(i)** We present the first comprehensive survey on knowledge distillation for vision-language models, providing a structured overview of methods, training objectives, and datasets (*who did what*) along with a comparative analysis of performance profiles, empirical trends, and trade-offs (*what works best and when*) . **(ii)** We propose a multi-view taxonomy that organizes VLM distillation along orthogonal axes, enabling practitioners to locate, compare, and select approaches suited to their specific settings. **(iii)** We identify open challenges and future directions that we hope will guide subsequent research in KD for VLMs. **Our anonymous repository is available at: Link.**

2 Background

2.1 Vision-Language Models

A vision-language model (VLM) integrates a visual encoder with an autoregressive language model to enable tasks that require joint visual-linguistic understanding. The dominant architecture follows an encoder-projector-LLM pipeline: a visual encoder f_v extracts patch-level representations from an input image x_v, a learned projection W_{proj} maps these representations into the language model's embedding space as visual tokens z_v:

$$z_v = W_{\text{proj}} \cdot f_v(x_v) \in \mathbb{R}^{N_v \times d} \tag{1}$$

where N_v is the number of visual tokens and d is the language embedding dimension. The projected tokens z_v are concatenated with the text embeddings to form a mixed-modality input sequence. Given a text instruction x_t, a VLM generates a response $y = (y_1, \ldots, y_L)$ autoregressively:

$$p(y \mid x_v, x_t) = \prod_{l=1}^{L} p(y_l \mid y_{<l}, z_v, x_t) \tag{2}$$

This architecture is shared across the majority of current VLMs, despite differences in encoder scale, projection design, and language backbone. It also underlies both full-scale models and and distilled compressed variants (Wang et al., 2020, 2023b; Zhu et al., 2023a; Shu et al., 2024).

2.2 Knowledge Distillation for VLMs

When knowledge distillation is applied to VLMs, distillation objective operates over the multimodally conditioned distributions defined in Eq. 2. Given a teacher VLM π_T and a student VLM π_S, each with their own visual encoder and projection, the token-level distillation objective minimizes:

$$\mathscr{L}_{\text{distill}} = \sum_{l=1}^{L} D\Big(\pi_T(\cdot \mid y_{<l}, z_v^T, x_t) \,\|\, \pi_S(\cdot \mid y_{<l}, z_v^S, x_t)\Big) \tag{3}$$

where D is a divergence measure and z_v^T, z_v^S are the teacher's and student's projected visual representations (Eq. 1). A critical property of Eq. 3 is that teacher and student condition on *different* visual evidence ($z_v^T \neq z_v^S$ in general), even for the same input image. This asymmetry, absent in text-only distillation, is the source of difficulties specific to the multimodal setting. We formalize this distinction by comparing Eq. 3 to its LLM counterpart and **derive the resulting challenges in Appendix A.**

3 A Taxonomy for Knowledge Distillation in Vision-Language Models

We organize VLM distillation methods along three orthogonal views, each addressing a distinct design question a practitioner faces when designing or selecting a distillation approach:

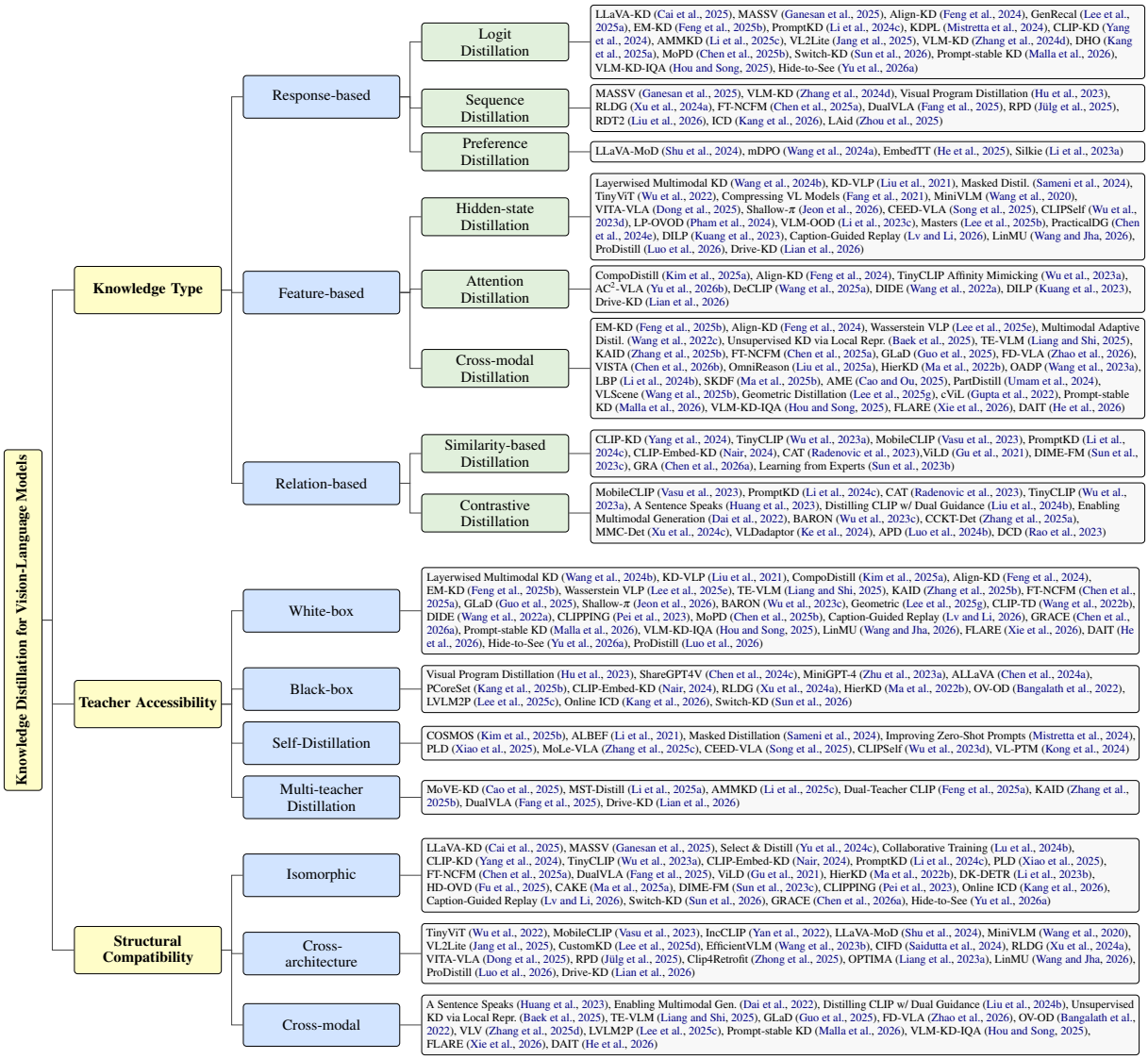

Figure 1: Taxonomy of VLM distillation methods.

- **Knowledge Type** asks *What does the student learn from the teacher?*
- **Teacher Accessibility** asks *What can be observed about the teacher?*
- **Structural Compatibility** asks *What must be bridged before distillation can proceed?*

While these axes co-constrain in practice–for instance, white-box access is required for feature distillation, and structural mismatches restrict baseline objective choices–co-constraint does not imply dependence. Two design axes remain independent if fixing a value along one does not uniquely dictate the choice on another. This independence holds across all axes: white-box isomorphic pairs admit logit, hidden-state, attention, or relation-based distillation, while black-box configurations support either sequence or preference transfer. Similarly, cross-architecture structures tolerate either black-box or white-box regimes. **The axes are thus orthogonal, following a natural design ordering**: practitioners typically assess structural compatibility first, determine teacher accessibility second, and select the specific knowledge type third. Moreover, to provide a clearer and more detailed overview of the literature, we summarize several major branches of our taxonomy into structured comparison tables in the appendix. **More detailed comparisons of representative methods can be found in Appendix I.** We now describe each view in detail.

3.1 Knowledge Type

Knowledge type axis categorizes *what* student learns from teacher. In VLMs, all teacher out-

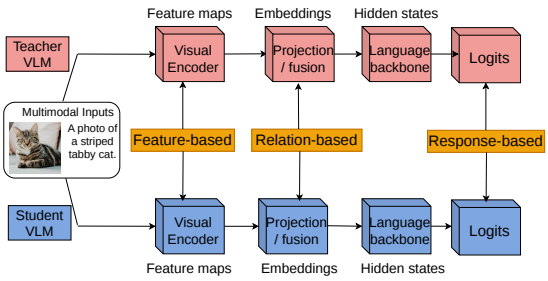

Figure 2: Knowledge type view of VLM distillation.

puts are multimodally conditioned: the teacher's response to any text prefix depends on its visual representation. As illustrated in Figure 2, VLM distillation objective (Eq. 3) can be instantiated in several ways depending on which aspect of teacher's behavior is transferred.

Response-based Distillation. Response-based methods distill the teacher's output-level behavior, varying in the granularity and form of the supervision signal. *Logit distillation* directly minimizes the per-step divergence between teacher and student next-token distributions:

$$\mathscr{L}_{\text{logit}} = \frac{1}{|y|}\sum_{t=1}^{|y|} \mathrm{KL}\big(p_t^T \,\|\, p_t^S\big) \tag{4}$$

where $p_t^T = p^T(y_t \mid y_{<t}, v, q)$ and $p_t^S = p^S(y_t \mid y_{<t}, v, q)$ denote the teacher and student output distributions conditioned on visual embedding v and text prefix q. This preserves the teacher's full output distribution including uncertainty and ranked alternatives at each decoding step. *Sequence distillation* replaces the soft distribution in Eq. 4 with discrete teacher-generated sequences $\hat{y} \sim p^T(\cdot \mid v, q)$, and trains the student with standard cross-entropy:

$$\mathscr{L}_{\text{seq}} = -\frac{1}{|\hat{y}|}\sum_{t=1}^{|\hat{y}|} \log p^S\big(\hat{y}_t \mid \hat{y}_{<t}, v, q\big) \tag{5}$$

By materializing the teacher's knowledge as text, this objective discards distributional information but enables black-box distillation where only generated outputs are available. *Preference distillation* casts teacher as a reference policy and training the student to distinguish preferred from dis-preferred responses via DPO objective (Rafailov et al., 2023):

$$\mathscr{L}_{\text{DPO}} = -\log \sigma\Big(\beta \log \tfrac{\pi_S(y_w|v,q)}{\pi_T(y_w|v,q)} - \beta \log \tfrac{\pi_S(y_l|v,q)}{\pi_T(y_l|v,q)}\Big) \tag{6}$$

where y_w and y_l are teacher-scored preferred and dis-preferred responses for the same (v, q) pair. This transfers the teacher's relative quality judgments rather than its token-level distributions.

Feature-based Distillation. Feature-based methods require white-box access and supervise the student's internal representations rather than its outputs. The cross-modal interface is the most important and uniquely VLM-specific target within this category. Three sub-types target different internal components: *hidden-state distillation* enforces layer-to-layer alignment of representations across the visual encoder, projection, and language backbone; *attention distillation* transfers the teacher's attention patterns, particularly cross-modal attention maps that encode which visual regions ground each text token; and *cross-modal feature distillation* operates directly on the projection-layer representations where visual and linguistic information are fused, targeting the modality interface. Methods of this type have also been applied to segmentation tasks, where segmentation networks are trained to align their internal representations with pretrained VLM features (Ma et al., 2022a; Liang et al., 2023b; Zhou et al., 2023; Ding et al., 2022).

Relation-based Distillation. Relation-based methods transfer the structural relationships among the teacher's embeddings rather than its individual predictions. *Similarity-based distillation* penalizes discrepancies in the teacher's $b \times b$ pairwise similarity matrix across a mini-batch of embeddings $\{h_i\}$:

$$\mathscr{L}_{\text{sim}} = \big\|S^T - S^S\big\|_F^2, \quad S_{ij} = \frac{h_i^\top h_j}{\|h_i\|\|h_j\|} \tag{7}$$

This captures inter-sample relationships—which images are semantically close, which image-text pairs are well-aligned—that logit matching cannot express. *Contrastive distillation* goes further by training the student to reproduce the teacher's complete batch similarity distribution, capturing positive alignments and the near-miss structure among hard negatives rather than only the scalar similarity values. Objectives of both types have been applied beyond compression to transfer the relational structure of vision-language embedding spaces into segmentation networks (Xu et al., 2022; Ma et al., 2022a; Zhou et al., 2023).

3.2 Teacher Accessibility

The accessibility axis determines which knowledge signals are observable and thus constrains

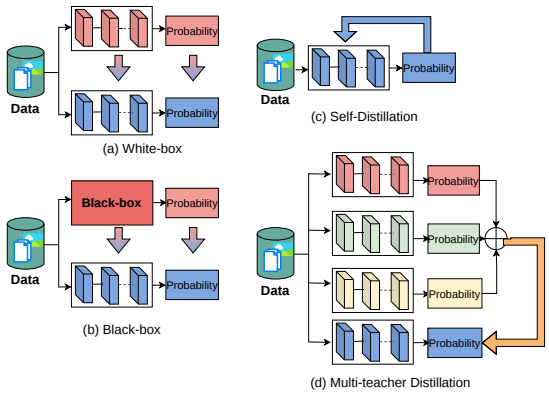

Figure 3: Teacher accessibility view of VLM distillation.

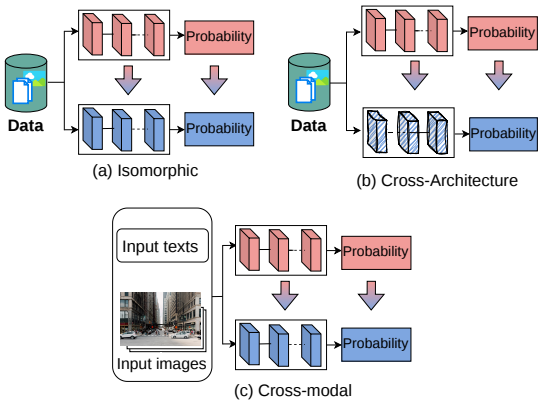

Figure 4: Structural compatibility view of VLM distillation.

which knowledge types from Section 3.1 are feasible (Figure 3). *White-box distillation* grants access to teacher's internal representations, including weights, hidden states, attention maps, and intermediate activations. Such access is required for feature-based and attention-based knowledge types. *Black-box distillation* restricts student to the teacher's observable outputs, typically generated text. *Self-distillation* uses no external teacher and instead derives supervision from the model's own alternative forward passes or cross-modal consistency constraints. *Multi-teacher distillation* combines signals from multiple specialized teachers, each capturing a complementary aspect of visual or linguistic information. This regime requires coordinating heterogeneous supervision signals and managing capacity allocation across teachers.

3.3 Structural Compatibility

The structural compatibility axis characterizes what must be bridged before any distillation objective can be applied (Figure 4). *Isomorphic* pairs share the same vocabulary, architecture, and modality configuration, differing only in scale. No bridging is required: all knowledge types are directly applicable. *Cross-architecture* pairs share the same input-output modality structure but differ in internal architecture of the visual encoder, the language backbone, or both. Feature-level alignment requires intermediate projection layers to bridge representational differences. *Cross-modal* pairs involve a VLM teacher that jointly processes vision and language, and a student that operates in a fundamentally different paradigm. This is the most extreme structural mismatch, where student lacks the architectural components needed to replicate teacher's cross-modal processing. Cross-modal knowledge transfer from VLMs to segmentation models has also been explored in both 2D and 3D settings (Ma et al., 2022a; Zhou et al., 2023; Xu et al., 2022; Umam et al., 2024).

While this survey focuses on general VLM compression, our taxonomy also extends to task specific architectures. For example, **vision-language embedding models** (Appendix F) mainly rely on relation based and contrastive distillation for global ranking, while **dense segmentation networks** (Appendix G) and **open-vocabulary object detectors** (Appendix H) highlight the challenge of transferring global image text semantics to pixel or region level predictions and **vision-language-action models** (Appendix E) leverage distillation to transfer privileged sensing, high-level reasoning into embodied agents. Together, these applications show how the same distillation principles adapt to different architectures and tasks.

4 Distillation for Vision-Language Models

We next review how specific methods instantiate these dimensions. Along with a descriptive summary, we further **examine the comparative performance trade-offs, empirical trends, and fragilities that govern these systems in Appendix C.**

4.1 Knowledge Types

4.1.1 Response-based Distillation.

Logit distillation. Switch-KD (Sun et al., 2026) is a visual-switch distillation framework that unifies vision-language knowledge transfer within a shared text-probability space. LLaVA-KD (Cai et al., 2025) extends Eq. (4) with modality-separated KL terms over visual and language tokens (MDist) and

a cosine similarity regulariser over the visual token self-correlation matrix (RDist). Align-KD (Feng et al., 2024) replaces the forward KL with its reverse form $\mathrm{KL}(p^S \| p^T)$ to encourage mean-seeking optimisation, combined with an attention-map distillation term on the first decoder layer. VLsI (Lee et al., 2024) generalises to every intermediate layer by projecting hidden states into vocabulary space via lightweight verbaliser heads.

Sequence distillation. By materializing the teacher's knowledge as discrete text, sequence distillation provides supervision for student models across a variety of applications, such as building self-distilled fine-tuning sets (Ganesan et al., 2025), guiding action learning via hindsight relabeling (Sumers et al., 2023b), filtering valid candidate programs for reasoning (Hu et al., 2023), supplying sequence-level semantic supervision for vision encoders (Zhang et al., 2024d), and dynamically generating sparse demonstrations to enhance student models via online in-context learning (Kang et al., 2026).

Preference distillation. LLaVA-MoD (Shu et al., 2024) applies Eq. (6) in a two-stage protocol: mimic distillation via KL logit matching is followed by preference distillation where the large MLLM serves as π_T. mDPO (Wang et al., 2024a) prevents models from ignoring images in multimodal DPO by introducing an image-conditioned term and a reward anchor to maintain the chosen response's likelihood.

4.1.2 Feature-based Distillation.

Hidden-state distillation. Providing internal representational supervision, this approach guides student models by enforcing layer-wise cross-component alignment (Wang et al., 2024c), grounding tokens via object-detection teachers (Liu et al., 2021), reconstructing masked regions with teacher hidden states (Sameni et al., 2024), transferring semantics via soft labels (Wu et al., 2022), and regressing on teacher states to build linear-complexity architectures (Wang and Jha, 2026).

Attention distillation. Explicitly transferring cross-modal attention patterns provides targeted supervision to resolve compositional reasoning failures via direct attention matching (Kim et al., 2025a), enable efficient mobile deployment through dynamic token selection (Feng et al., 2024), transfer sequential perception and reasoning for autonomous driving from multiple teachers (Lian et al., 2026), and facilitate action-context-aware adaptive computation for low-latency robotics (Yu et al., 2026b).

Cross-modal feature distillation. Direct alignment at the modality interface shapes student representations through mechanisms such as reweighting informative cross-modal signals (Feng et al., 2025b; Wang et al., 2022c; Liang and Shi, 2025), matching joint distributions via optimal transport to capture global geometry (Lee et al., 2025e), and anchoring patch-level visual features for unsupervised local transfer (Baek et al., 2025).

4.1.3 Relation-based Distillation.

Similarity-based distillation. CLIP-KD (Yang et al., 2024) proposes a Relation Distillation (RD) component that aligns pairwise image-text similarity distributions via KL divergence, and further combines it with Contrastive Representation Distillation (CRD) to treat instance pairs as positives and negatives for relational matching. COSMOS (Kim et al., 2025b) applies cross-modal self-distillation by aligning similarity distributions between global and local multi-modal crops through an EMA teacher, preserving cross-modal relational structure without external supervision.

Contrastive distillation. Instead of matching individual predictions, contrastive distillation preserves the teacher's cross-modal similarity geometry to provide relational supervision for student models across a variety of applications, such as matching batch-level image–text structures (Vasu et al., 2023; Wu et al., 2023b), transferring fine-grained intra-sample visual token relationships via centered kernel alignment (Chen et al., 2026a), enforcing relational consistency across different prompting contexts (Li et al., 2024c), enhancing alignment quality through hard-negative mining (Radenovic et al., 2023), and adapting to specialized domains like generative transfer and structured human-body representations (Huang et al., 2023; Liu et al., 2024b; Dai et al., 2022).

4.2 Accessibility

4.2.1 White-box Distillation

Access to intermediate hidden states is the most common form, supporting layerwise feature matching and compression via internal alignment. LMD (Wang et al., 2024b), KD-VLP (Liu et al., 2021), SF-CLIP (Sameni et al., 2024), TinyViT (Wu et al., 2023b), Compressing VLMs via KD (Fang et al., 2021), and MiniVLM (Wang et al., 2020) all supervise student with teacher activations from intermediate encoder or trans-

former layers, enabling component-level alignment, masked-feature reconstruction, attention-head supervision, or stage-wise layer-reduction distillation that cannot be recovered from outputs alone.

Access to attention weight matrices enables explicit transfer of the teacher's internal reasoning patterns. CompoDistill (Kim et al., 2025a) and Align-KD (Feng et al., 2024) match teacher cross-attention or self-attention maps to preserve compositional grounding and spatial focus, while TinyCLIP (Wu et al., 2023b) transfers the teacher's cross-modal affinity structure through similarity-matrix matching.

Access to cross-modal projection outputs enables direct supervision of image–text alignment at the representation boundary through mechanisms such as matching teacher projection-layer embeddings (Feng et al., 2025b; Lee et al., 2025e; Wang et al., 2022c), establishing patch-level correspondences (Baek et al., 2025), and exploiting richer internal signals like transfer entropy and hidden-state trajectories to dynamically guide knowledge transfer (Liang and Shi, 2025; Zhang et al., 2025b).

Access to output logit distributions remains the least structural but still richer than hard labels. LLaVA-KD (Cai et al., 2025) and GenRecal (Lee et al., 2025a) distill next-token teacher distributions through token-level KL alignment, preserving uncertainty and ranked alternatives, though this requires the teacher to remain loaded during training for joint forward passes.

4.2.2 Black-box Distillation

VPD (Hu et al., 2023) generates teacher's structured visual programs over a question corpus, and only these sequences serve as student supervision. ShareGPT4V (Chen et al., 2024c) constructs a large-scale caption dataset by prompting GPT-4V for high-quality image descriptions, then uses these captions as sequence-level supervision targets for open-source VLM students. MiniGPT-4 (Zhu et al., 2023a) and ALLaVA (Chen et al., 2024a) adopt the same approach, using GPT-3.5/GPT-4 generated instruction-response pairs as black-box supervision for instruction-following distillation.

4.2.3 Self-Distillation

COSMOS (Kim et al., 2025b) generates cross-modal supervision entirely from the model itself via bidirectional image-text consistency, and remarkably matches external-teacher distillation on VLP benchmarks at substantially lower training cost. Before this, ALBEF (Li et al., 2021) uses a momentum encoder as a self-distillation target to stabilise VLP contrastive learning, validating that alignment before cross-attention fusion prevents spurious correlations. In a related but prompt-centric direction, KDPL (Mistretta et al., 2024) regularizes learned prompts against domain-specific overfit by distilling consistency between teacher's prompted and unprompted predictions on a diverse unlabeled set.

4.2.4 Multi-teacher Distillation

MoVE-KD (Cao et al., 2025) proposes a Mixture-of-LoRA-Experts student with a gating network that dynamically routes inputs to the expert specialised for each teacher encoder (CLIP, SAM, DINOv2). Extending the multi-teacher paradigm from visual encoders to task-specialised VLMs, MST-Distill (Li et al., 2025a) employs multiple task-specialised VLM teachers with dynamic confidence-based weighting to capture complementary capabilities across reasoning, counting, spatial comprehension, and open-domain VQA. To avoid capacity fragmentation under multiple teachers, AMMKD (Li et al., 2025c) applies curriculum-style adaptive weighting that down-weights teachers the student has already approached, directing limited capacity toward the most productive signal. In a more targeted dual-teacher setup, LongD-CLIP (Feng et al., 2025a) uses specialised short-text and long-text CLIP teachers to address CLIP's short-text training bias while extending context length.

4.3 Structural Compatibility

4.3.1 Isomorphic

Isomorphic distillation compresses models within matched architectural families and vocabularies. In the generative domain, this enables direct scale reduction across identical LLaVA frameworks (Cai et al., 2025; Ganesan et al., 2025) and fixed-family VLMs (Yu et al., 2024c; Lu et al., 2024b). GRACE (Chen et al., 2026a) advances this setting by unifying isomorphic transfer with quantization-aware training, allowing same-family students to preserve full-precision visual structures even under aggressive low-bit constraints. Similarly, scale-matched contrastive compression relies on standard CLIP ViTs that share the exact same text tokenization (Yang et al., 2024; Wu et al., 2023b; Nair, 2024; Li et al., 2024c; Radenovic et al., 2023; Wei et al., 2023; Vasu et al., 2023; Liu et al., 2024b).

4.3.2 Cross-Architecture

Cross-architecture distillation transfers knowledge across diverse topologies through mechanisms such as establishing cross-scale recipes from deep to shallow ViTs (Wu et al., 2022), adapting ViT teachers into lightweight CNNs for mobile deployment (Vasu et al., 2023), mitigating drift in continual learning via generative replay (Yan et al., 2022), distilling dense backbones into sparse Mixture-of-Experts (Shu et al., 2024), designing compact models with reduced multimodal components (Wang et al., 2020), leveraging importance-based distillation for custom architectures (Jang et al., 2025; Lee et al., 2025d), reducing transformer layers (Fang et al., 2021), and preserving key information across architectural boundaries using rate–distortion modules (Saidutta et al., 2024).

4.3.3 Cross-Modal

Knowledge transfer across fundamentally different architectural boundaries allows students to inherit semantic grounding without shared structures (Xue et al., 2022), facilitating applications like adapting multimodal representations into unimodal visual students (Huang et al., 2023), shifting from discriminative to generative models (Dai et al., 2022), specializing visual encoders for domain-specific body understanding (Liu et al., 2024b), and employing trainable intermediate teachers to adaptively filter and transfer frozen VLM representations into lightweight classifiers for fine-grained categorization (He et al., 2026). Adapting to specialized domains (detailed in Appendix E – H), this paradigm transfers image-level semantics to region proposals for open-vocabulary object detection (Gu et al., 2021; Bangalath et al., 2022), extracting dense pseudo-labels for segmentation (Zhou et al., 2022, 2023), distilling RL trajectories and 3D geometric priors into vision-language-action models (Xu et al., 2024a; Guo et al., 2025), and leveraging MLLM judges to supervise lightweight embedding models (Li et al., 2026; Gu et al., 2025b)

5 Open Challenges and Future Directions

While VLM distillation is expanding rapidly, critical bottlenecks still limit current VLM compression paradigms. Overcoming these open challenges outlines key future directions for the field, which we analyze in greater depth in Appendix A.

▷ ***Unified Cross-Modal Distillation Objectives.*** Challenges identified in Appendix A.2 are currently handled by separate loss terms. Whether these components reinforce or interfere with each other when composed remains unknown. Framing VLM distillation as a structured alignment problem over joint visual-linguistic token space could replace ad hoc objective design with a more principled framework.

▷ ***Tokenization Heterogeneity Across Modalities.*** Cross-tokenizer distillation is well studied in LLMs, but GenRecal (Lee et al., 2025a) remains the only VLM method addressing mismatched text vocabularies. VLM setting compounds this: teacher and student may simultaneously differ in text vocabulary, visual patch size, and visual token count.

▷ ***Cross-Modal Transfer Under Black-Box Access.*** Feature-based methods that most effectively transfer visual grounding require white-box access. Under black-box access, VLM distillation remains limited to sequence-level transfer through teacher-generated text, where grounding is only implicit. Recovering cross-modal structure from output sequences alone would enable distillation from proprietary multimodal models, but requires fundamentally new objective designs.

▷ ***Adaptation to Evolving Architectures.*** Most surveyed methods target encoder-projector-LLM pipeline, where projection layer serves as a natural cross-modal distillation target. This assumption breaks down for native multimodal architectures, mixture-of-experts VLMs, and models for video or 3D inputs. Distillation objectives must therefore extend to architectures in which the cross-modal interface is distributed rather than localized.

6 Conclusion

This paper has presented the first dedicated survey of knowledge distillation for vision-language models. The multi-view taxonomy proposed in this work organizes VLM distillation methods along knowledge type, teacher accessibility, and structural compatibility perspectives, accompanied by a systematic compilation of training and evaluation datasets across VLM distillation settings. Our survey reveals that addressing VLM distillation goes beyond adapting text-only objectives: it requires unified cross-modal alignment, structural bridging across heterogeneous tokenizers, and evaluation protocols that measure grounding fidelity rather than aggregate accuracy. We hope this survey provides a structured reference and a solid foundation for future research on KD for VLMs.

Limitations

While this survey provides the first dedicated taxonomy for knowledge distillation in vision-language models, it has limitations. Distillation methods for video-language and 3D vision-language models may not be fully represented. Due to the rapid development of the field, the proposed taxonomy will require timely updates to incorporate new methods and paradigms. We hope that this survey serves as a foundation for both methodological and empirical advances in VLM distillation.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*.

Changwoo Baek, Jou Won Song, and Kyeongbo Kong. 2025. Unsupervised knowledge distillation via local representations for vision-language models. *IEEE Signal Processing Letters*, 32:3147–3151.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of CVPR*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. 2025. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 239–249.

Guiming Cao and Yuming Ou. 2025. Ame: Aligned manifold entropy for robust vision-language distillation. *ArXiv*, abs/2508.08644.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. 2025. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19846–19856.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note about kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *arXiv preprint arXiv:2102.08981*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.

Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Mohamed Elhoseiny, and Sean Chang Culatana. 2023a. Exploring open-vocabulary semantic segmentation without human labels. *Preprint*, arXiv:2306.00450.

Kewei Chen, Yayu Long, Shuai Li, and Mingsheng Shang. 2025a. Ft-ncfm: An influence-aware data distillation framework for efficient vla models. *arXiv preprint arXiv:2511.16233*.

Lin Chen, Jaya Cui, Bowen Zhang, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

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

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, and 2 others. 2023b. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*.

Yang Chen, Shuai Fu, and Yu Zhang. 2025b. Mopd: Mixture-of-prompts distillation for vision-language models. *IEEE Transactions on Multimedia*.

Yanlong Chen, Amirhossein Habibian, Luca Benini, and Yawei Li. 2026a. Gated relational alignment via confidence-based distillation for efficient vlms. *Preprint*, arXiv:2601.22709.

Yiye Chen, Yanan Jian, Xiaoyi Dong, Shuxin Cao, Jing Wu, Patricio Vela, Benjamin E Lundell, and Dongdong Chen. 2026b. Vista: Enhancing visual conditioning via track-following preference optimization in vision-language-action models. *arXiv preprint arXiv:2602.05049*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Hongying Meng. 2024e. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23501–23511.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional map of the world. In *Proceedings of CVPR*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of CVPR*.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of AISTATS*.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of CVPR*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024a. Ultrafeedback: Boosting language models with scaled ai feedback. In *Proceedings of ACL*.

Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. 2024b. Sinkhorn distance minimization for knowledge distillation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14846–14858.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. 2025. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23724–23732.

Dengxin Dai and Wen Yang. 2011. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and Remote Sensing Letters*, 8(1):173–176.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. In *Findings*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. *Preprint*, arXiv:2112.07910.

Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, and 1 others. 2025. Vita-vla: Efficiently teaching vision-language models to act via action expert distillation. *arXiv preprint arXiv:2510.09607*.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, and 1 others. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, and 1 others. 2026. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1):17.

Zhen Fang, Zhuoyang Liu, Jiaming Liu, Hao Chen, Yu Zeng, Shiting Huang, Zehui Chen, Lin Chen, Shanghang Zhang, and Feng Zhao. 2025. Dualvla: Building a generalizable embodied agent via partial decoupling of reasoning and action. *arXiv preprint arXiv:2511.22134*.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1408–1418.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. 2024. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. *ArXiv*, abs/2412.01282.

Yuheng Feng, Changsong Wen, Zelin Peng, Li Jiaye, and Siyu Zhu. 2025a. Retaining knowledge and enhancing long-text representations in clip through dual-teacher distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24895–24904.

Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. 2025b. Em-kd: Distilling efficient multimodal large language model with unbalanced vision tokens. *ArXiv*, abs/2511.21106.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Shenghao Fu, Junkai Yan, Qize Yang, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. 2025. A hierarchical semantic distillation framework for open-vocabulary object detection. *IEEE Transactions on Multimedia*.

Mugilan Ganesan, Shane Segal, Ankur Aggarwal, Nish Sinnadurai, Sean Lie, and Vithursan Thangarasa. 2025. Massv: Multimodal adaptation and self-data distillation for speculative decoding of vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.

Yuying Ge, Keqiang Sun, Junting Pan, Yibin Qian, and 1 others. 2024. Seed-data-edit technical report: A hybrid dataset for instruction-guided image editing. *arXiv preprint arXiv:2405.04007*.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of CVPR*.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. 2022. Scaling open-vocabulary image segmentation with image-level labels. *Preprint*, arXiv:2112.12143.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, and 1 others. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *NeurIPS Datasets and Benchmarks Track*.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025a. Breaking the modality barrier: Universal embedding learning with multimodal llms. *Preprint*, arXiv:2504.17432.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. 2025b. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *Preprint*, arXiv:2510.13515.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

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

Minghao Guo, Meng Cao, Jiachen Tao, Rongtao Xu, Yan Yan, Xiaodan Liang, Ivan Laptev, and Xiaojun Chang. 2025. Glad: Geometric latent distillation for vision-language-action models. *arXiv preprint arXiv:2512.09619*.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of CVPR*.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2022. cvil: Cross-lingual training of vision-language models using knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1734–1741. IEEE.

Danna Gurari, Qing Li, Alan J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of CVPR*.

Eric He, Akash Gupta, Adian Liusie, Vatsal Raina, Piotr Molenda, Shirom Chabra, and Vyas Raina. 2025. Embedding the teacher: Distilling vllm preferences for scalable image retrieval. *ArXiv*, abs/2510.12014.

Zhengxu He, Jun Li, and Zhijian Wu. 2026. Dait: Distillation from vision-language models to lightweight classifiers with adaptive intermediate teacher transfer. *Preprint*, arXiv:2603.15166.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Yongkang Hou and Jiarun Song. 2025. Visual-language model knowledge distillation method for image quality assessment. *Preprint*, arXiv:2507.15680.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9590–9601.

Zeyi Huang, Andy Zhou, Zijia Lin, Mu Cai, Haohan Wang, and Yong Jae Lee. 2023. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11651–11661.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of CVPR*.

Jinseong Jang, Chunfei Ma, and Byeongwon Lee. 2025. Vl2lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30073–30083.

Boseong Jeon, Yunho Choi, and Taehan Kim. 2026. Shallow-{\pi}: Knowledge distillation for flow-based vlas. *arXiv preprint arXiv:2601.20262*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Peng-Tao Jiang and Yuqi Yang. 2023. Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. *Preprint*, arXiv:2305.01275.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. 2024. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of CVPR*.

Tobias Jülg, Wolfram Burgard, and Florian Walter. 2025. Refined policy distillation: From vla generalists to rl experts. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11677–11684. IEEE.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of ICCV*.

Seongjae Kang, Dong Bok Lee, Hyungjoon Jang, and Sung Ju Hwang. 2025a. Simple yet effective semi-supervised knowledge distillation from vision-language models via dual-head optimization.

Seongjae Kang, Dong Bok Lee, Hyungjoon Jang, Dongseop Kim, and Sung Ju Hwang. 2025b. Pcoreset: Effective active learning through knowledge distillation from vision-language models. *arXiv preprint arXiv:2506.00910*.

Zhiqi Kang, Rahaf Aljundi, Vaggelis Dorovatas, and Karteek Alahari. 2026. Online in-context distillation for low-resource vision language models. *Preprint*, arXiv:2510.18117.

Junjie Ke, Lihuo He, Bo Han, Jie Li, Di Wang, and Xinbo Gao. 2024. Vldadaptor: Domain adaptive object detection with vision-language model distillation. *IEEE Transactions on Multimedia*, 26:11316–11331.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *arXiv preprint arXiv:1603.07396*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Chris Fitzpatrick, Peter Bull, Greg Lipstein, Julian Nelli, Ron Zhu, and 1 others. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. 2025a. Compodistill: Attention distillation for compositional reasoning in multimodal llms. *ArXiv*, abs/2510.12184.

Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. 2025b. Cosmos: Cross-modality self-distillation for vision language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14690–14700.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*.

Jun Kong, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2024. Multimodality self-distillation for fast inference of vision and language pretrained models. *IEEE Transactions on Multimedia*, 26:8928–8940.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of ICCV Workshops*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Huafeng Kuang, Jie Wu, Xiawu Zheng, Ming Li, Xuefeng Xiao, Rui Wang, Min Zheng, and Rongrong Ji. 2023. Dlip: Distilling language-image pre-training. *arXiv preprint arXiv:2308.12956*.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of CVPR*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Byung-Kwan Lee, Ryo Hachiuma, Yonghyun Ro, Yu-Chiang Frank Wang, and Yueh-Hua Wu. 2025a. Genrecal: Generation after recalibration from large to small vision-language models. *ArXiv*, abs/2506.15681.

Byung-Kwan Lee, Ryo Hachiuma, Yu-Chiang Frank Wang, Yonghyun Ro, and Yueh-Hua Wu. 2024. Vlsi: Verbalized layers-to-interactions from large to small vision language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29545–29557.

Byung-Kwan Lee, Yu-Chiang Frank Wang, and Ryo Hachiuma. 2025b. Masking teacher and reinforcing student for distilling vision-language models. *arXiv preprint arXiv:2512.22238*.

Donghoon Lee, Tung M Luu, Younghwan Lee, and Chang D Yoo. 2025c. Sample efficient reinforcement learning via large vision language model distillation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. 2025d. Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25176–25186.

Lung-Hao Lee, Hsiu-Min Shih, and Jin Wang. 2025e. Modality-specific Knowledge Distillation with Wasserstein Distance Minimization for Vision-Language Pretrained Models . *IEEE Transactions on Artificial Intelligence*, 1(01):1–13.

Lung-Hao Lee, Hsiu-Min Shih, and Jin Wang. 2025f. Modality-specific knowledge distillation with wasserstein distance minimization for vision-language pretrained models. *IEEE Transactions on Artificial Intelligence*.

Seonho Lee, Jiho Choi, Inha Kang, Jiwook Kim, Junsung Park, and Hyunjung Shim. 2025g. 3d-aware vision-language models fine-tuning with geometric distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10628–10647.

Bo Li, Yuanhan Zhang, Chunyuan Li, Tianle Zhang, Yuhang Lyu, Kai Zhang, Haoyi Dong, Yuchen Huang, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022a. Language-driven semantic segmentation. *Preprint*, arXiv:2201.03546.

Hui Li, Pengfei Yang, Juanyang Chen, Le Dong, Yanxin Chen, and Quan Wang. 2025a. Mst-distill: Mixture of specialized teachers for cross-modal knowledge distillation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1588–1597.

Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. 2024b. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16678–16687.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *ArXiv*, abs/2312.10665.

Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. 2023b. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6501–6510.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022b. Grounded language-image pre-training. In *Proceedings of CVPR*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *Preprint*, arXiv:2601.04720.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. 2023c. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2492–2503.

Yifan Li, Yifan Du, Kun Zhang, Di Zhou, Jinpeng Wang, and 1 others. 2023d. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*.

Yiming Li, Yi Wang, Wenqian Wang, Dan Lin, Bingbing Li, and Kim-Hui Yap. 2025b. Open world object detection: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):988–1008.

Yuqi Li, Chuanguang Yang, Junhao Dong, Zhengtao Yao, Haoyan Xu, Zeyu Dong, Hansheng Zeng, Zhulin An, and Yingli Tian. 2025c. Ammkd: Adaptive multimodal multi-teacher distillation for lightweight vision-language models. *arXiv preprint arXiv:2509.00039*.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. 2024c. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626.

Zuolei Li, Xingyu Gao, Xiaofan Wang, and Jianlong Fu. 2025d. Latbot: Distilling universal latent actions for vision-language-action models. *arXiv preprint arXiv:2511.23034*.

Weitong Lian, Zecong Tang, Haoran Li, Tianjian Gao, Yifei Wang, Zixu Wang, Lingyi Meng, Tengju Ru, Zhejun Cui, Yichen Zhu, Hangshuo Cao, Qi Kang, Tianxing Chen, Yusen Qin, Kaixuan Wang, and Yu Zhang. 2026. Drive-kd: Multi-teacher distillation for vlms in autonomous driving. *Preprint*, arXiv:2601.21288.

Chen Liang, Jiahui Yu, Ming-Hsuan Yang, Matthew Brown, Yin Cui, Tuo Zhao, Boqing Gong, and Tianyi Zhou. 2023a. Module-wise adaptive distillation for multimodality foundation models. *Advances in Neural Information Processing Systems*, 36:69719–69735.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Xinyuan Song, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. 2023b. Open-vocabulary semantic segmentation with mask-adapted clip. *Preprint*, arXiv:2210.04150.

Jessica E. Liang and Jianbo Shi. 2025. TE-VLM: Transfer entropy for vision language model distillation.

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

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *Preprint*, arXiv:2212.09506.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024a. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.

Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. 2024b. Distilling clip with dual guidance for learning discriminative human body shape representation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 256–266.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of CVPR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Pei Liu, Qingtian Ning, Xinyan Lu, Haipeng Liu, Weiliang Ma, Dangen She, Peng Jia, Xianpeng Lang, and Jun Ma. 2025a. Omnireason: A temporal-guided vision-language-action framework for autonomous driving. *arXiv preprint arXiv:2509.00789*.

Shilong Liu and 1 others. 2025b. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Songming Liu, Bangguo Li, Kai Ma, Lingxuan Wu, Hengkai Tan, Xiao Ouyang, Hang Su, and Jun Zhu. 2026. Rdt2: Exploring the scaling limit of umi data towards zero-shot cross-embodiment generalization. *arXiv preprint arXiv:2602.03310*.

Yifan Liu, Yifan Du, Yifan Zhang, Yu Li, Yao Feng, Wen Chen, and 1 others. 2024d. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. 2021. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. In *NAACL-HLT*.

Yuliang Liu, Weicong Li, Biao Yang, and 1 others. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Shichen Lu, Longteng Guo, Wenxuan Wang, Zijia Zhao, Tongtian Yue, Jing Liu, and Si Liu. 2024b. Collaborative training of tiny-large vision language models. In *ACM Multimedia 2024*.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2018. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. 2023. Plug-and-play, dense-label-free extraction of open-vocabulary semantic segmentation from vision-language models. *CoRR*, abs/2311.17095.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and Yansheng Li. 2024a. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*.

Lin Luo, Xin Wang, Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. 2024b. Adversarial prompt distillation for vision-language models. *arXiv preprint arXiv:2411.15244*.

Siqi Luo, Chunlei Meng, Hongda Zhang, Zhongxue Gan, and Chun Ouyang. 2026. Prodistill: A progressive prompting framework for fine-grained vlm distillation. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13282–13286.

Jiaming Lv and Peihua Li. 2026. Distilling knowledge from caption-guided replay for VLM-based continual learning.

Timo Lüddecke and Alexander S. Ecker. 2022. Image segmentation using text and image prompts. *Preprint*, arXiv:2112.10003.

Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2022a. Open-vocabulary semantic segmentation with frozen vision-language models. *Preprint*, arXiv:2210.15138.

Shiyuan Ma, Donglin Qian, Kai Ye, and Shengchuan Zhang. 2025a. Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 5982–5990.

Shuailei Ma, Yuefeng Wang, Ying Wei, Enming Zhang, Jiaqi Fan, Xinyu Sun, and Peihao Chen. 2025b. Skdf: a simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. 2022b. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14074–14083.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report, arXiv preprint arXiv:1306.5151.

Hiranya Jeet Malla, Milad Bazli, and Mehrdad Arashpour. 2026. Prompt-stable knowledge distillation of vision-language models for efficient waste classification in material recovery facilities. *Resources, Conservation and Recycling*, 230:108885.

Amir M Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, Alireza Taherian, Kimia Dinashi, Amirali Miri, and 1 others. 2025. A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of CVPR*.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*.

Anand Mishra, Karteek Alahari, and C. V. Jawahar. 2012. Scene text recognition using higher order language priors. *Proceedings of the British Machine Vision Conference*.

Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D Bagdanov. 2024. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *European conference on computer vision*, pages 459–477. Springer.

Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. 2024. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18:100605.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of CVPR*.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*.

Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, and Giovanni Semeraro. 2025. xvlm2vec: Adapting lvlm-based embedding models to multilinguality using self-knowledge distillation. *Preprint*, arXiv:2503.09313.

Lakshmi Nair. 2024. Clip-embed-kd: Computationally efficient knowledge distillation using embeddings as teachers. *arXiv preprint arXiv:2404.06170*.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*.

Chao Pang, Yi Zhao, Yang Liu, and 1 others. 2024. Vhm: Versatile and honest vision language model for remote sensing image analysis. *arXiv preprint arXiv:2403.20213*.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *Proceedings of CVPR*.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18983–18992.

Chau Pham, Truong Vu, and Khoi Nguyen. 2024. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 779–788.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*.

Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, and Xingang Wang. 2023. Freeseg: Unified, universal and open-vocabulary image segmentation. *Preprint*, arXiv:2303.17225.

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. *Proceedings of CITS*.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash J. Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Kumar Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6967–6977.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jun Rao, Liang Ding, Shuhan Qi, Meng Fang, Yang Liu, Li Shen, and Dacheng Tao. 2023. Dynamic contrastive distillation for image-text retrieval. *IEEE Transactions on Multimedia*, 25:8383–8395.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of EMNLP*.

Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Jaejin Cho, Ching-Hua Lee, Chouchang Yang, Yilin Shen, and Hongxia Jin. 2024. CIFD: Controlled information flow to enhance knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sepehr Sameni, Kushal Kafle, Hao Tan, and Simon Jenni. 2024. Building Vision-Language Models on Solid Foundations with Masked Distillation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14216–14226, Los Alamitos, CA, USA. IEEE Computer Society.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Siva Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. 2025. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(3):e70036.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, Haoyuan Li, Bolin Li, Zhelun Yu, Si Liu, Hongsheng Li, and Hao Jiang. 2024. Llava-mod: Making llava tiny via moe knowledge distillation. *ArXiv*, abs/2408.15881.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, Marcus Rohrbach, and Dhruv Batra. 2019. Towards vqa models that can read. In *Proceedings of CVPR*.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Yuxin Huang, Han Zhao, Donglin Wang, and Haoang Li. 2025. Ceed-vla: Consistency vision-language-action model with early-exit decoding. *arXiv preprint arXiv:2506.13725*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of SIGIR*.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*.

Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. 2023a. Distilling internet-scale vision-language models into embodied agents. *arXiv preprint arXiv:2301.12507*.

Theodore R. Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. 2023b. Distilling internet-scale vision-language models into embodied agents. In *International Conference on Machine Learning*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting the unreasonable effectiveness of data in deep learning era. In *Proceedings of ICCV*.

Haoyi Sun, Xiaoxiao Wang, Ning Mao, Qian Wang, Lifu Mu, Wen Zheng, Tao Wei, and Wei Chen. 2026. Switch-kd: Visual-switch knowledge distillation for vision-language models. *Preprint*, arXiv:2604.14629.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. 2023a. Journeydb: A benchmark for generative image understanding. *arXiv preprint arXiv:2307.00716*.

Lina Sun, Yewen Li, and Yumin Dong. 2023b. Learning from expert: Vision-language knowledge distillation for unsupervised cross-modal hashing retrieval. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. 2023c. Dime-fm: Distilling multimodal and efficient foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15521–15533.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023d. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. 2024. Partdistill: 3d shape part segmentation by vision-language model distillation. *Preprint*, arXiv:2312.04016.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. 2023. Mobileclip: Fast image-text models through multi-modal reinforced training. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15963–15974.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention*.

Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M. Patel, and Ran Xu. 2023. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. *Preprint*, arXiv:2303.16891.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *ArXiv*, abs/2406.11839.

Hongjie Wang and Niraj K. Jha. 2026. Linmu: Multimodal understanding made linear. *Preprint*, arXiv:2601.01322.

Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, L. Zhang, Jianfeng Gao, and Zicheng Liu. 2020. Minivlm: A smaller and faster vision-language model. *ArXiv*, abs/2012.06946.

Jin Wang, Dawei Liao, You Zhang, Dan Xu, and Xuejie Zhang. 2024b. Layerwised multimodal knowledge distillation for vision-language pretrained model. *Neural Networks*, 175:106272.

Jin Wang, Dawei Liao, You Zhang, Dan Xu, and Xuejie Zhang. 2024c. Layerwised multimodal knowledge distillation for vision-language pretrained model. *Neural Networks*, 175:106272.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. 2025a. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14824–14834.

Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. 2023a. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11186–11196.

Meng Wang, Huilong Pi, Ruihui Li, Yunchuan Qin, Zhuo Tang, and Kenli Li. 2025b. Vlscene: Vision-language guidance distillation for camera-based 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7808–7816.

Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. 2023b. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13899–13913.

Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2022a. Distilled dual-encoder model for vision-language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8901–8913.

Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. 2022b. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*.

Zhecan Wang, Noel C. F. Codella, Yen-Chun Chen, Luowei Zhou, Xiyang Dai, Bin Xiao, Jianwei Yang, Haoxuan You, Kai-Wei Chang, Shih-Fu Chang, and Lu Yuan. 2022c. Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks. *ArXiv*, abs/2204.10496.

Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. 2024d. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of AAAI*.

Yixuan Wei, Han Hu, Zhenda Xie, Ze Liu, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. 2023. Improving clip fine-tuning performance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5416–5426.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, and 1 others. 2024a. Towards open vocabulary learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):5092–5113.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Chen, Xinggang Wang, Hongyang Chao, and Han Hu. 2023a. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21913–21923.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, and 1 others. 2023b. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. 2022. Tinyvit: Fast pretraining distillation for small vision transformers. *ArXiv*, abs/2207.10666.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023c. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15254–15264.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. 2023d. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. 2024b. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *Preprint*, arXiv:2310.01403.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2025. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024c. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baizhaoliang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. Aid: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. 2010. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium: 100 Years ISPRS*.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of CVPR*.

Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Yuqi Xie, Fengyuan Hu, Jimmy Wu, Zhengyi Luo, Linxi Fan, and 1 others. 2025. Self-improving vision-language-action models with data generation via residual rl. *arXiv preprint arXiv:2511.00091*.

Zihao Xiao, Longlong Jing, Shangxuan Wu, Alex Zihao Zhu, Jingwei Ji, Chiyu Max Jiang, Wei-Chih Hung, Thomas Funkhouser, Weicheng Kuo, Anelia Angelova, and 1 others. 2024. 3d open-vocabulary panoptic segmentation with 2d-3d vision-language distillation. In *European Conference on Computer Vision*, pages 21–38. Springer.

Chengen Xie, Chonghao Sima, Tianyu Li, Bin Sun, Junjie Wu, Zhihui Hao, and Hongyang Li. 2026. Flare: Learning future-aware latent representations from vision-language models for autonomous driving. *Preprint*, arXiv:2601.05611.

Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. 2024a. Rldg: Robotic generalist policy distillation via reinforcement learning. *arXiv preprint arXiv:2412.09858*.

Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Luntong Li, and Yonghong Tian. 2025. Vlms-guided representation distillation for efficient vision-based reinforcement learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29534–29544.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. *Preprint*, arXiv:2112.14757.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024b. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yifan Xu, Mengdan Zhang, Xiaoshan Yang, and Changsheng Xu. 2024c. Exploring multi-modal contextual knowledge for open-vocabulary object detection. *IEEE Transactions on Image Processing*, 33:6253–6267.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.

Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. 2022. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, pages 22–38. Springer.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025a. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–27.

Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. 2025b. Clip-cid: Efficient clip distillation via cluster-instance discrimination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21974–21982.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Seonghoon Yu, Dongjun Nam, Byung-Kwan Lee, and Jeany Son. 2026a. Hide to see: Reasoning-prefix masking for visual-anchored thinking in vlm distillation. *Preprint*, arXiv:2605.11651.

Tianyu Yu, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Wenda Yu, Tianshi Wang, Fengling Li, Jingjing Li, and Lei Zhu. 2026b. Acˆ 2-vla: Action-context-aware adaptive computation in vision-language-action models for efficient robotic manipulation. *arXiv preprint arXiv:2601.19634*.

Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. 2024c. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, pages 219–236. Springer.

Xi Yue, Huan Zhang, Licheng Zheng, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*.

Chuhan Zhang, Chaoyang Zhu, Pingcheng Dong, Long Chen, and Dong Zhang. 2025a. Cyclic contrastive knowledge transfer for open-vocabulary object detection. *arXiv preprint arXiv:2503.11005*.

Da Zhang, Feiyu Wang, Bingyu Li, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. 2025b. Kaid: Knowledge-aware interactive distillation for vision-language models. pages 3212–3221.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024b. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. 2025c. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. *arXiv preprint arXiv:2503.20384*.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. 2024c. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18164–18181.

Tiezheng Zhang, Yitong Li, Yu-cheng Chou, Jieneng Chen, Alan Yuille, Chen Wei, and Junfei Xiao. 2025d. Vision-language-vision auto-encoder: Scalable knowledge distillation from diffusion models. *arXiv preprint arXiv:2507.07104*.

Zaiwei Zhang, Gregory P. Meyer, Zhichao Lu, Ashish Shrivastava, Avinash Ravichandran, and Eric M. Wolff. 2024d. Vlm-kd: Knowledge distillation from vlm for long-tail visual recognition. *ArXiv*, abs/2408.16930.

Bei Zhao, Yanfei Zhong, Liangpei Zhang, and Bo Huang. 2016. The fisher kernel coding framework for high spatial resolution scene classification. *Remote Sensing*, 8(2):157.

Ruiteng Zhao, Wenshuo Wang, Yicheng Ma, Xiaocong Li, Francis EH Tay, Marcelo H Ang Jr, and Haiyue Zhu. 2026. Fd-vla: Force-distilled vision-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2602.02142*.

Li Zhong, Ahmed Ghazal, Jun-Jun Wan, Frederik Zilly, Patrick Mackens, Joachim Vollrath, and Bogdan Coseriu. 2025. Clip4retrofit: Enabling real-time image labeling on edge devices via cross-architecture clip distillation. In *Proceedings of the computer vision and pattern recognition conference*, pages 3829–3837.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of CVPR*.

Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract free dense labels from clip. *Preprint*, arXiv:2112.01071.

Haoyi Zhou, Shuo Li, Tianyu Chen, Qi Song, Chonghan Gao, and Jianxin Li. 2025. Towards long-window anchoring in vision-language model distillation. *arXiv preprint arXiv:2512.21576*.

Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. 2023. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Preprint*, arXiv:2212.03588.

Bryan Zhu, Nicholas Lui, Jeremy Irvin, Jimmy Le, Sahil Tadwalkar, Chenghao Wang, Zutao Ouyang, Frankie Y. Liu, Andrew Y. Ng, and Robert B. Jackson. 2022. Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping. *arXiv preprint arXiv:2207.11166*.

Chaoyang Zhu and Long Chen. 2024. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8954–8975.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, and Chunhua Shen. 2023b. Segprompt: Boosting open-world segmentation via category-level prompt learning. *Preprint*, arXiv:2308.06531.

A LLMs–VLMs Connection, Challenges and Open Directions

This appendix reviews the LLM distillation framework that VLM distillation inherits, then identifies the challenges that arise when the visual modality is introduced.

A.1 LLM Distillation Objective

Knowledge distillation (Hinton et al., 2015) trains a compact student to approximate a stronger teacher using supervision signals richer than hard labels: output distributions, generated sequences, intermediate hidden states, attention patterns, or preference rankings. These knowledge types, together with the level of access to the teacher (white-box vs. black-box (Yang et al., 2025a)) and structural constraints such as cross-tokenizer alignment (Zhang et al., 2024c; Cui et al., 2025), define the main design axes of distillation methods. For autoregressive LLMs, the standard objective (Hinton et al., 2015; Kim and Rush, 2016) minimizes a per-token divergence over a sequence x:

$$\mathscr{L}_{\text{distill}}^{\text{LLM}} = \sum_{t=1}^{|x|} D(\pi_T(\cdot \mid x_{<t}) \,\|\, \pi_S(\cdot \mid x_{<t})) \quad (8)$$

The divergence D has been instantiated as forward KL (Hinton et al., 2015; Kim and Rush, 2016), reverse KL (Gu et al., 2023; Wu et al., 2025), Jensen-Shannon divergence (Wen et al., 2023; Agarwal et al., 2024), skewed variants (Ko et al., 2024), and Wasserstein distances (Cui et al., 2024b; Lee et al., 2025f). The context $x_{<t}$ may be drawn from teacher generations (off-policy) or the student's own outputs (on-policy) (Agarwal et al., 2024). These text-only methods are now well surveyed (Xu et al., 2024b; Yang et al., 2025a; Fang et al., 2026). A key property of Eq. 8 is that both teacher and student condition on exactly the same input: the text prefix $x_{<t}$.

A.2 VLM-Specific Distillation Challenges and Open Directions

Comparing the VLM distillation objective (Eq. 3, Section 2.2) to its LLM counterpart (Eq. 8) reveals a fundamental difference: the conditioning no longer matches. In the LLM case, both models see the same text prefix $x_{<t}$. In the VLM case, each model produces its own visual representation through a separate encoder f_v and projection W_{proj}, so $z_v^T \neq z_v^S$ in general. Teacher and student therefore see different visual evidence even for the same input image. This asymmetry introduces four challenges absent in text-only distillation.

Unified Cross-Modal Distillation Objectives. Because the teacher's token distribution depends on z_v^T, a student whose z_v^S is weaker or differently structured cannot match the teacher's outputs regardless of the language backbone's capacity. Minimizing Eq. 3 directly does not correct a poor z_v^S; the visual encoder and projection must be explicitly supervised to reduce the gap between z_v^T and z_v^S. Currently, these difficulties are handled by method-specific loss terms derived independently, such as LLaVA-KD (Cai et al., 2025) (introducing visual token self-correlation regularizers with a modality-separated KL term), Align-KD (Feng et al., 2024) (targeting decoder attention maps), or EM-KD (Feng et al., 2025b) (reweighting tokens). What remains unresolved is whether these components reinforce or destructively interfere when composed on the joint space, making a single structured alignment functional a key open problem.

Tokenization Heterogeneity Across Modalities. The projection W_{proj} and the self-attention layers where visual and text tokens interact are the most VLM-specific components, yet Eq. 3 supervises only the final output distribution and provides no gradient signal targeted at these components. Hidden-state matching treats all layers uniformly without distinguishing cross-modal coupling from unimodal processing. This gap is severely compounded by tokenization heterogeneity across modalities. While LLMs handle vocabulary mismatch via embedding distribution alignment, VLMs are strictly harder because the teacher and student can simultaneously differ in text vocabulary, visual patch size, and visual token count N_v. GenRecal (Lee et al., 2025a) is the only method explicitly addressing text vocabularies but omits the visual tokenization dimension; no existing method resolves this dual boundary mismatch jointly.

Cross-Modal Transfer Under Black-Box Access. Eq. 3 sums over L language tokens, but the visual evidence supporting each token is spread across N_v visual tokens, where typically $N_v \gg L$. Uniform matching over visual positions is therefore dominated by low-information tokens (background, redundant patches) rather than the sparse regions that actually determine generation quality. EM-KD (Feng et al., 2025b) addresses this credit-assignment problem under white-box ac-

cess by reweighting cross-modal supervision toward linguistically attended visual tokens. However, this challenge becomes highly acute under black-box access to frontier APIs, where access is restricted to sequence-level transfer via teacher-generated text (Chen et al., 2024c; Hu et al., 2023), discarding internal distributional geometry. Recovering cross-modal structural signals from output sequences alone—inferring which image regions ground which text tokens without observing internal attention layers—is an information-theoretically nontrivial inversion problem requiring entirely new objective designs.

Adaptation to Evolving Architectures. The overwhelming majority of surveyed approaches assume an encoder-projector-LLM pipeline where the projection layer W_{proj} serves as a localized cross-modal distillation target. This assumption breaks down for at least three emerging architectural paradigms:

- **Native Multimodal Architectures:** Fusing vision and language without a discrete projection stage eliminates the primary supervision target of hidden feature distillation methods, making layer-to-layer alignment strategies ill-defined when the interface is distributed across interleaved attention layers.
- **Mixture-of-Experts (MoE) VLMs:** Models with dynamic per-token routing invalidate fixed layer correspondence assumptions, as the expert activated in the teacher does not necessarily map to a fixed layer in the student, rendering hidden-state matching ambiguous.
- **Variable-Length Contexts:** Multi-modal models for video or 3D inputs violate the fixed-N_v assumption. As the number of visual tokens varies dynamically across inputs, the cross-modal credit-assignment problem becomes highly fluid and can no longer be resolved by static spatial reweighting vectors.

B Literature Survey Methodology

To guarantee a systematic review, we established a paper collection and filtering protocol. This protocol outlines our search infrastructure, query expressions, venue constraints, and explicit eligibility criteria.

B.1 Search Architecture and Query Expressions

We systematically queried five primary indexing databases (the ACL Anthology, arXiv, Semantic Scholar, IEEE Xplore, and the ACM Digital Library). The search queries targeted titles, abstracts, and keywords with the following core search strings—applied both individually and in combination to cover the full breadth of the field:

- `"knowledge distillation"`, `"KD"`, `"affinity mimicking"`, `"representation transfer"`
- `"vision-language model"`, `"VLM"`, `"multimodal LLM"`, `"CLIP"`, `"vision-language-action"`, `"VLA"`
- `"compression"`, `"alignment"`, `"efficient"`, `"open-vocabulary"`

B.2 Venue Scope and Temporal Boundaries

Our curation pipeline spanned top-tier peer-reviewed venues, restricted to papers published between **January 2020 and May 2026**:

- Natural Language Processing: ACL, EMNLP, NAACL, COLING.
- Machine Learning & AI: NeurIPS, ICLR, ICML, AAAI.
- Computer Vision & Multimedia: CVPR, ICCV, ECCV, ACM MM, ICASSP, SIGIR.
- Core Journals: IEEE TPAMI, IJCV, JMLR, and various IEEE Transactions.

Non-peer-reviewed preprints from arXiv within this timeline were included to preserve coverage of urgent, recent work not yet appearing in archival proceedings.

B.3 Inclusion and Exclusion Criteria

Candidate papers were passed through strict eligibility filters:

- **Inclusion Criteria:** We included papers whose primary contribution involves the knowledge distillation of or from a model that jointly processes vision and language inputs.
- **Exclusion Criteria:** We excluded papers whose use of distillation is incidental (e.g., distillation as a minor training trick within a broader multimodal system).

B.4 Quantitative Distribution and Trajectory

Our protocol yielded a final pool of **114 distinct papers**. The chronological distribution of these selected works reveals an exponential growth curve that mirrors the macroeconomic shifts in multimodal AI research (Table 1).

Time Frame	Paper Count	Percentage (%)
2020 – 2021	5	4.5%
2022 – 2023	28	24.5%
2024 – May 2026	81	71%
Total	**114**	**100.0%**

Table 1: Chronological distribution of the surveyed VLM distillation papers.

This distribution directly reflects the growth trajectory of the field: VLM distillation emerged as a research area only after the widespread release of CLIP (2021) and large instruction-tuned VLMs, and has accelerated sharply since 2024, driven by the deployment of billion-parameter multimodal models.

C Comparative Analysis

To complement the categorical descriptions in Section 4, this section provides a critical comparison of empirical trends, trade-offs, and fragilities across representative state-of-the-art works. Performance benchmarks are systematized across the four primary VLM distillation settings to analyze when specific paradigms are preferable.

Generative VLM Distillation. Across recent generative VLM distillation work, a consistent convergence toward combining logit-level supervision with cross-modal feature distillation at the projection layer has emerged (Table 2). This convergence reflects a structural insight: logit distillation alone cannot compensate for a weak visual encoder because the teacher and student condition on different visual evidence (as formalized in Appendix A.2). Methods that address this grounding gap explicitly by supervising the cross-modal interface consistently outperform those relying on a single objective.

As shown in Table 2, LLaVA-KD and Switch-KD – both combining logit and cross-modal feature distillation – outperform LLaVA-MoD on MMB and GQA at comparable model sizes, indicating stronger visual grounding and instruction-following in the distilled student. Recent formulations like Switch-KD advance this strategy by unifying cross-modal supervision within a shared text-probability space. When paired preference data is available, preference distillation is particularly useful for hallucination-sensitive applications, as demonstrated by LLaVA-MoD achieving a strong POPE score. However, such comparisons remain confounded by architecture, backbone, and training data variations. Under aggressive compression, trends suggest that adaptive balancing of distillation components becomes critical for smaller student regimes.

Contrastive VLM Distillation. A field-wide shift from similarity-based toward contrastive distillation has occurred in CLIP compression (Table 3). This shift is driven by a fundamental property of CLIP's downstream utility: zero-shot classification and retrieval depend strictly on inter-sample relational structure rather than absolute embedding magnitudes, which contrastive distillation preserves explicitly.

In Table 3, MobileCLIP2-S0 (74.8M parameters) achieves 71.5 on ImageNet and 43.7 on COCO T→I, outperforming TinyCLIP-61M (115.4M parameters). This comparison suggests that objective design can outweigh raw parameter counts, though the result remains partially confounded by MobileCLIP2's utilization of stronger teachers and higher-quality training data. Overall, contrastive distillation establishes a robust default for CLIP-style compression when preserving zero-shot classification and cross-modal retrieval mechanics.

Open-Vocabulary Object Detection. The literature undergoes a measurable progression in transfer granularity: image-level transfer (ViLD) → bag-of-regions (BARON) → region-level feature distillation (CORA+) → cyclic contrastive transfer (CCKT-Det++). This trajectory is driven by a fundamental mismatch: CLIP models are trained on global image-text pairs and their features are not natively localizable, making image-level transfer increasingly lossy since object detection requires fine-grained region discrimination.

Each architectural shift addresses this localization gap at a finer granularity, but Table 4 also reveals a clear trade-off: CCKT-Det++ substantially improves novel-category accuracy ($AP_N^{50} = 46.0$), while ViLD retains higher overall base accuracy ($AP^{50} = 51.3$). This suggests that finer-grained contrastive transfer is preferable for open-vocabulary generalization, whereas coarser image-level trans-

Method	Distillation Type	MMB	VQA	GQA	POP	TVQA
MoVE-KD	Multi-teacher + Feature	67.4	79.9	63.9	86.3	59.6
MoVE-KD	Multi-teacher + Feature	48.8	73.8	57.7	86.1	44.3
LLaVA-MoD	Logit + Preference	66.3	75.8	58.7	87.0	58.5
LLaVA-KD	Logit + Feature	71.0	80.3	62.5	86.7	59.7
LLaVA-KD	Logit + Feature	61.3	77.7	59.8	86.4	52.0
Switch-KD	Logit + Feature	71.4	81.4	63.9	86.8	60.3
Switch-KD	Logit + Feature	63.5	79.6	61.6	87.3	52.3

Table 2: Generative VLM Distillation benchmarks across representative language backbones.

Method	Distillation Type	Params	IN	Flickr	COCO
TinyCLIP-40M	Attention + Similarity	85.2M	59.8	59.1	33.5
TinyCLIP-61M	Attention + Similarity	115.4M	62.4	62.6	36.5
MobileCLIP-S0	Contrastive	53.8M	67.8	67.7	40.4
MobileCLIP-S1	Contrastive	84.9M	72.6	71.0	44.0
MobileCLIP-S2	Contrastive	99.1M	74.4	73.4	45.4
MobileCLIP2-S0	Contrastive	74.8M	71.5	69.2	43.7
MobileCLIP2-S2	Contrastive	99.1M	77.2	74.8	48.8
MobileCLIP2-B	Contrastive	149.7M	79.4	76.5	49.9

Table 3: Contrastive VLM Distillation benchmarks on ImageNet validation and text-to-image retrieval.

fer may better preserve overall base detection balance.

Vision-Language Embedding Distillation. Under a unified B3-Qwen2-2B teacher and FastVLM-0.5B student baseline (Table 5), relational and contrastive objectives like RKD and CKD lead on classification tasks. This indicates they better preserve the global semantic structure necessary for recognition tasks where inter-sample relationships matter most.

In contrast, logit combined with cross-modal feature distillation (EM-KD) leads on VQA tasks, reflecting stronger transfer of fine-grained reasoning tied directly to visual token supervision. Notably, standard MSE-based pointwise feature alignment achieves the highest OK-VQA score (52.63), indicating that direct representation proximity remains highly competitive when downstream tasks reward vector proximity rather than relational structure. Across all settings, objective selection must be driven explicitly by the downstream task requirements.

Observations. Synthesizing results across these distinct sub-domains yields three overarching insights:

1. **Relational superiority:** Relational and contrastive objectives consistently outperform purely pointwise feature matching when downstream success depends on preserving ranking, retrieval, or global cross-modal alignment topologies.

2. **Dynamic composition behavior:** While hybrid, multi-signal objectives generally improve baseline performance, they are not universally sufficient; the relative importance and interaction of each individual component shift drastically under severe student model compression.

3. **Fragility of architectural homogeneity:** Teacher-student architectural homogeneity remains the most fragile assumption in the literature. Layer-level alignments require strictly compatible internal hidden dimensions and break down under cross-architecture deployment. When cross-architecture compatibility is uncertain, contrastive and logit-based objectives provide significantly more portability, whereas feature-level supervision should be reserved for architecturally identical model pairs.

D Datasets and Evaluations

This section summarises the datasets commonly used for training and evaluating VLM distillation methods.

D.1 Datasets

D.1.1 Datasets for Training

As summarized in Table 6, the data landscape for vision–language model (VLM) training has

Method	Distillation Type	OV-COCO AP_N^{50}	OV-COCO AP^{50}
ViLD	Feature	27.6	51.3
BARON	Contrastive	34.0	53.5
CORA	Feature	35.1	35.4
CORA+	Feature	41.7	43.8
CCKT-Det	Contrastive	38.0	35.0
CCKT-Det++	Contrastive	45.3	46.9
$\text{CCKT-Det++}_{\text{Swin}}$	Contrastive	46.0	46.2

Table 4: Open-Vocabulary Object Detection results on novel (AP_N^{50}) and base classes.

Method	Distillation Type	ImageNet	N24News	VOC2007	CLS Avg	OK-VQA	DocVQA	I-VQA	VQA Avg
MSE	Feature-based	52.90	70.17	78.03	64.92	52.63	74.20	34.03	51.45
RKD	Similarity-based	53.73	70.40	79.80	66.17	51.13	73.93	35.10	51.32
CKD	Contrastive	53.13	70.47	78.47	66.01	51.87	74.83	35.00	51.38
EMO	Sim. + Feature	52.60	68.27	80.03	63.84	51.03	73.30	34.03	50.66
EM-KD	Logit + Feature	53.20	67.23	78.23	64.24	51.17	77.17	36.80	51.56

Table 5: Vision-Language Embedding Model Distillation performance profiles under a unified baseline.

evolved from relatively small, curated English-caption datasets toward web-scale and increasingly multilingual corpora. Early public resources such as SBU Caption (Ordonez et al., 2011), COCO Caption (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and Conceptual Captions (Sharma et al., 2018) established the basic image–text supervision paradigm, whereas later datasets such as WIT, LAION400M, and LAION5B substantially expanded both scale and linguistic coverage. At the same time, several influential pretrained VLMs were built on large non-public corpora, such as CLIP, ALIGN, FILIP, and WebLI, indicating that proprietary data remains a major factor in scaling high-capacity models. Another clear trend is the growing separation between *pre-training* data and *post-training* data: while representation learning typically relies on massive web-scale corpora, knowledge distillation and instruction tuning increasingly depend on smaller but more curated datasets, including LLaVA-style instruction data, preference-learning resources, and edit-oriented or feedback-based supervision. Domain-specific adaptation follows a similar pattern. In remote sensing, for example, recent works combine large-scale alignment corpora with comparatively smaller instruction-tuning datasets, reflecting a broader shift from generic cross-modal alignment toward task-aware and domain-specialized supervision (Table 6).

D.1.2 Datasets for Evaluation

Table 7–11 show that VLM evaluation has become increasingly broad and multi-dimensional, extending well beyond conventional image classification or retrieval. Classical vision benchmarks such as ImageNet, CIFAR, COCO, and VOC remain important anchors for measuring core perceptual performance, but recent evaluation practice places much greater emphasis on multimodal reasoning, visual question answering, hallucination robustness, OCR capability, and instruction-following behavior. This trend is reflected in the growing use of specialized benchmarks such as MME, MMBench, MMMU, MMStar, HallusionBench, MathVista, OCRBench, and MMVet, which probe more fine-grained capabilities than traditional recognition datasets alone. At the same time, evaluation is becoming increasingly domain-aware: remote-sensing VLMs are now assessed not only on classification and retrieval, but also on visual question answering, grounding, and broader knowledge-oriented benchmarks. Overall, the benchmark distribution in Table 7–11 suggests that modern VLM evaluation is no longer centered on a single task family; instead, it requires a diverse suite of datasets that jointly capture perception, reasoning, robustness, and domain generalization.

D.2 Evaluation Metrics in Vision-Language Model Distillation

Evaluation in vision-language model (VLM) distillation is typically performed using standard downstream task metrics rather than distillation-specific criteria. In practice, the most commonly reported metrics are accuracy-based measures for classification and question answering, ranking-based measures for retrieval, and overlap-based measures for

Table 6: Summary of datasets used for VLM training and distillation. † Num. of samples refers to instruction/preference pairs rather than raw image-text pairs. ‡ Auxiliary dataset; may not consist solely of image-text pairs. § Num. of samples refers to region descriptions rather than images.

Dataset	Year	Image-Text Pairs	Language	Public
General VLM Pre-training (Public)				
SBU Caption (Ordonez et al., 2011)	2011	1M	English	✓
COCO Caption (Lin et al., 2014)	2015	1.5M	English	✓
YFCC100M (Thomee et al., 2016)	2015	100M	English	✓
Visual Genome (VG) (Krishna et al., 2017)	2017	5.4M§	English	✓
Conceptual Captions (CC3M) (Sharma et al., 2018)	2018	3.3M	English	✓
Localized Narratives (LN) (Pont-Tuset et al., 2020)	2020	0.87M	English	✓
Conceptual 12M (CC12M) (Changpinyo et al., 2021)	2021	12M	English	✓
Wikipedia-based Image Text (WIT) (Srinivasan et al., 2021)	2021	37.6M	108 Languages	✓
Red Caps (RC) (Desai et al., 2021)	2021	12M	English	✓
LAION400M (Schuhmann et al., 2021)	2021	400M	English	✓
LAION5B (Schuhmann et al., 2022)	2022	5B	> 100 Languages	✓
WuKong (Gu et al., 2022)	2022	100M	Chinese	✓
General VLM Pre-training (Non-public)				
CLIP (Radford et al., 2021)	2021	400M	English	×
ALIGN (Jia et al., 2021)	2021	1.8B	English	×
FILIP (Yao et al., 2021)	2021	300M	English	×
WebLI (Chen et al., 2023b)	2022	10B	109 Languages	×
Auxiliary / Supplementary Pre-training				
JFT3B‡ (Sun et al., 2017)	—	∼3B images	English	×
C4‡ (Raffel et al., 2020)	2020	∼750 GB text	English	✓
Object365‡ (Shao et al., 2019)	2019	638K images	English	✓
Gold-G‡ (Kamath et al., 2021)	2021	0.8M	English	✓
VLM Knowledge Distillation & Instruction Tuning				
RLAIF-V-Dataset† (Yu et al., 2024b)	2024	83K	English	✓
Mantis-Instruct† (Jiang et al., 2024)	2024	721K	English	✓
LLaVA-558K† (Liu et al., 2023b)	2023	558K	English	✓
LLaVA-mix-665K† (Liu et al., 2024c)	2023	665K	English	✓
LLaVA-OneVision† (Li et al., 2024a)	2024	∼4M	English	✓
JourneyDB (Sun et al., 2023a)	2023	4M+	English	✓
SEED-Data-Edit (Part-2&3) (Ge et al., 2024)	2024	147K	English	✓
UltraFeedback† (Cui et al., 2024a)	2023	64K	English	✓
RLHF-V† (Yu et al., 2024a)	2024	1.4K/5.7K (HF)	English	✓
Remote Sensing Domain				
LHRS-Align (Muhtar et al., 2024)	2024	1.15M	English	✓
GeoChat-Instruct† (Kuckreja et al., 2024)	2024	306K	English	✓
LHRS-Instruct† (Muhtar et al., 2024)	2024	39.8K	English	✓
SkySenseGPT† (Luo et al., 2024a)	2024	381K	English	✓
SkyScript (Wang et al., 2024d)	2024	2.6M / 5.2M	English	✓
RemoteCLIP (Liu et al., 2024a)	2024	—	English	✓
VHM (Pang et al., 2024)	2024	—	English	✓

Table 7: Summary of datasets used for VLM evaluation (Part I: image classification). "—" denotes not applicable or not reported.

Dataset	Year	Classes	Training	Testing	Eval. Metric
Image Classification					
MNIST (LeCun et al., 1998)	1998	10	60,000	10,000	Accuracy
Caltech-101 (Fei-Fei et al., 2004)	2004	102	3,060	6,085	Mean Per Class
PASCAL VOC 2007 Classif. (Everingham et al., 2010)	2007	20	5,011	4,952	11-point mAP
Oxford 102 Flowers (Nilsback and Zisserman, 2008)	2008	102	2,040	6,149	Mean Per Class
CIFAR-10 (Krizhevsky, 2009)	2009	10	50,000	10,000	Accuracy
CIFAR-100 (Krizhevsky, 2009)	2009	100	50,000	10,000	Accuracy
ImageNet-1K (Deng et al., 2009)	2009	1,000	1,281,167	50,000	Accuracy
SUN397 (Xiao et al., 2010)	2010	397	19,850	19,850	Accuracy
SVHN (Netzer et al., 2011)	2011	10	73,257	26,032	Accuracy
STL-10 (Coates et al., 2011)	2011	10	1,000	8,000	Accuracy
GTSRB (Stallkamp et al., 2011)	2011	43	26,640	12,630	Accuracy
KITTI Distance (Geiger et al., 2012; Zhai et al., 2019)	2012	4	6,770	711	Accuracy
IIIT5K (Mishra et al., 2012)	2012	36	2,000	3,000	Accuracy
Oxford-IIIT Pets (Parkhi et al., 2012)	2012	37	3,680	3,669	Mean Per Class
Stanford Cars (Krause et al., 2013)	2013	196	8,144	8,041	Accuracy
FGVC Aircraft (Maji et al., 2013)	2013	100	6,667	3,333	Mean Per Class
FER2013 (Goodfellow et al., 2015)	2013	8	32,140	3,574	Accuracy
Rendered SST2 (Zhai et al., 2019)	2013	2	7,792	1,821	Accuracy
Describable Textures (DTD) (Cimpoi et al., 2014)	2014	47	3,760	1,880	Accuracy
Food-101 (Bossard et al., 2014)	2014	102	75,750	25,250	Accuracy
Birdsnap (Berg et al., 2014)	2014	500	42,283	2,149	Accuracy
RESISC45 (Cheng et al., 2017)	2017	45	3,150	25,200	Accuracy
CLEVR Counts (Johnson et al., 2017)	2017	8	2,000	500	Accuracy
PatchCamelyon (Veeling et al., 2018)	2018	2	294,912	32,768	Accuracy
EuroSAT (Helber et al., 2019)	2019	10	10,000	5,000	Accuracy
Hateful Memes (Kiela et al., 2020)	2020	2	8,500	500	ROC AUC
Country211 (Radford et al., 2021)	2021	211	43,200	21,100	Accuracy

generation. For classification-style tasks, the most common metric is *Accuracy*, defined as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\},$$

where $\hat{y}_i$ and y_i denote the predicted and ground-truth labels for the i-th example. In some settings, especially when class imbalance is relevant, *F1-score* is also reported:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

These metrics are widely used when evaluating whether the distilled student preserves the teacher's decision quality on classification, VQA, or multiple-choice multimodal reasoning tasks. For retrieval-oriented VLMs, especially CLIP-style dual-encoder models, the dominant metrics are *Recall@K* and sometimes *Mean Reciprocal Rank (MRR)*. Recall@K measures whether the correct target appears in the top-K retrieved results:

$$\text{Recall@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\text{rank}_i \le K\},$$

where rank_i is the rank of the correct match for query i. MRR is defined as

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\text{rank}_i}.$$

Among these, R@1, R@5, and R@10 are the most frequently reported, with R@1 usually serving as the headline retrieval metric. For generative VLMs, such as captioning or instruction-following models, the most common automatic metrics are *BLEU*, *ROUGE-L*, and *CIDEr*. BLEU evaluates n-gram precision with a brevity penalty:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where p_n is the modified n-gram precision, w_n is the corresponding weight, and BP denotes the

Table 8: Summary of datasets used for VLM evaluation (Part II: retrieval, action recognition, detection, and segmentation). "—" denotes not applicable or not reported.

Dataset	Year	Classes	Training	Testing	Eval.Metric
Image-Text Retrieval					
Flickr30K (Young et al., 2014)	2014	—	31,783	—	Recall
COCO Caption (Lin et al., 2014)	2015	—	82,783	5,000	Recall
Action Recognition					
UCF101 (Soomro et al., 2012)	2012	101	9,537	1,794	Accuracy
Kinetics700 (Carreira et al., 2019)	2019	700	494,801	31,669	Mean (Top-1, Top-5)
RareAct (Miech et al., 2020)	2020	122	7,607	—	mWAP, mSAP
Object Detection					
COCO 2014 Detection (Lin et al., 2014)	2014	80	83,000	41,000	box mAP
COCO 2017 Detection (Lin et al., 2014)	2017	80	118,000	5,000	box mAP
LVIS (Gupta et al., 2019)	2019	1,203	118,000	5,000	box mAP
ODinW (Li et al., 2022b)	2022	314	132,413	20,070	box mAP
Semantic Segmentation					
PASCAL VOC 2012 Seg. (Everingham et al., 2010)	2012	20	1,464	1,449	mIoU
PASCAL Context (Mottaghi et al., 2014)	2014	459	4,998	5,105	mIoU
Cityscapes (Cordts et al., 2016)	2016	19	2,975	500	mIoU
ADE20K (Zhou et al., 2017)	2017	150	25,574	2,000	mIoU

brevity penalty. ROUGE-L is based on the longest common subsequence (LCS) between a candidate sequence Y and a reference sequence X:

$$R_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{|X|}, \qquad P_{\text{LCS}} = \frac{\text{LCS}(X,Y)}{|Y|},$$

with the corresponding F-score

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}.$$

CIDEr measures consensus between the generated caption and reference captions using TF-IDF weighted n-gram similarity:

$$\text{CIDEr} \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{\mathbf{g}_n^{\text{cand}} \cdot \mathbf{g}_n^{\text{ref}(m)}}{\|\mathbf{g}_n^{\text{cand}}\| \, \|\mathbf{g}_n^{\text{ref}(m)}\|},$$

where $\mathbf{g}_n$ denotes the TF-IDF representation of n-grams. In practice, CIDEr is often treated as one of the most informative single metrics for captioning-style evaluation.

Some works additionally report calibration quality using *Expected Calibration Error (ECE)*, which measures the discrepancy between model confidence and empirical correctness:

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

where B_m is the m-th confidence bin. Although less universal than task accuracy or retrieval metrics, ECE is useful when the goal is to assess whether distillation improves the reliability of the student's confidence estimates. Overall, the most commonly used metrics in VLM distillation are therefore *Accuracy/F1* for classification-style tasks, *Recall@K* (especially R@1, R@5, and R@10) and *MRR* for retrieval, and *BLEU*, *ROUGE-L*, and *CIDEr* for generation. Their choice is mainly determined by the downstream task on which the distilled student is evaluated.

D.3 Critical Gaps in Current Benchmarking Practices

While the downstream benchmarks listed in Table 7–11 provide valuable indicators of aggregate student capability, a critical analysis of current evaluation protocols reveals a fundamental limitation: *they measure absolute task accuracy rather than successful distillation*. That is, current metrics confirm whether a student model can solve a given task, but fail to verify whether the student has faithfully inherited the teacher's specific behavioral, representational, and architectural properties, or independently achieved similar performance via alternate shortcuts. We identify four critical gaps in current

Table 9: Summary of datasets used for VLM evaluation (Part III: VQA and general multimodal capability benchmarks). "—" denotes not applicable or not reported.

Dataset	Year	Classes	Training	Testing	Eval.Metric
Multimodal Visual Question Answering					
VQAv2 (Goyal et al., 2017)	2017	—	443,757	81,434	VQA Score
OK-VQA (Marino et al., 2019)	2019	—	9,009	5,046	Accuracy
TextVQA (Singh et al., 2019)	2019	—	34,602	5,734	Accuracy
GQA (Hudson and Manning, 2019)	2019	—	943,000	12,578	Accuracy
VizWiz (Gurari et al., 2018)	2020	—	20,523	8,000	Accuracy
ScienceQA (image) (Lu et al., 2022)	2022	—	12,726	4,241	Accuracy
Multimodal Capability Benchmarks					
MME (Fu et al., 2023)	2023	14	—	2,194	MME Score
MMBench (Liu et al., 2024d)	2023	—	—	4,329	Accuracy
MMBench-CN (Liu et al., 2024d)	2023	—	—	4,329	Accuracy
POPE (Li et al., 2023d)	2023	—	—	9,000	Accuracy / F1
MMMU (Yue et al., 2024)	2024	30	—	11,500	Accuracy
MMStar (Chen et al., 2024b)	2024	6	—	1,500	Accuracy
HallusionBench (Guan et al., 2024)	2024	—	—	1,129	Accuracy
MathVista (mini) (Lu et al., 2024a)	2024	—	—	1,000	Accuracy
OCRBench (Liu et al., 2023c)	2024	—	—	1,000	OCRBench score
AI2D (Kembhavi et al., 2016)	2022	—	12,413	3,088	Accuracy
MMVet (Yu et al., 2023)	2023	6	—	218	GPT-4 based scoring

VLM distillation evaluation practice.

Grounding Fidelity. Standard VQA benchmarks such as VQAv2 measure answer accuracy but not whether the student attends to the same visual regions as the teacher. A student can produce correct answers by exploiting language priors without visual grounding. To our knowledge, no existing VLM distillation paper directly measures grounding fidelity directly, for instance by comparing teacher-student attention map alignment on held-out inputs or by evaluating on grounding-specific benchmarks such as RefCOCO, Visual Genome Relation , or the ARO benchmark (Attribute, Relation, and Order). This gap is particularly consequential given that cross-modal feature distillation methods such as Align-KD and EM-KD are explicitly motivated by the goal of transferring visual grounding—yet their success is evaluated only through aggregate VQA accuracy.

Hallucination. POPE measures object hallucination via binary existence questions, and HallusionBench targets the entanglement between language priors and visual evidence, but these benchmarks are rarely used jointly in distillation evaluations. More importantly, no paper examines whether the student's hallucination profile matches the teacher's. This distinction matters: a distilled model that hallucinates less than the teacher has not faithfully reproduced it; one that hallucinates more has introduced a new failure mode not present in the teacher. Reporting hallucination metrics relative to the teacher baseline, using complementary tools such as Object HalBench and MMHal-Bench alongside POPE, would provide a more complete characterization of what distillation preserves or degrades.

Robustness Under Distribution Shifts. For contrastive VLMs, works such as MobileCLIP reports IN-shift - the average zero-shot accuracy across five ImageNet benchmarks: ImageNet-V2, ImageNet-A, ImageNet-O, ImageNet-R, and ObjectNet —alongside standard IN-val accuracy, enabling assessment of representational robustness. However, generative VLM distillation papers report neither of these robustness metrics. A student that matches teacher accuracy on in-distribution benchmarks such as MMB or VQAv2 but degrades faster under distribution shift has not successfully inherited the teacher's representational robustness. Adopting equivalent out-of-distribution evaluation protocols for generative VLMs would be a mean-

Table 10: Summary of datasets used for VLM evaluation (Part IV: hallucination, captioning, generation, and editing). "—" denotes not applicable or not reported. For GEdit-Bench, G_SC, G_PQ, and G_O denote the GPT-based Semantic Consistency, Perceptual Quality, and Overall scores, respectively. For GenEval, Overall Score denotes the mean correctness score across six compositional generation tasks.

Dataset	Year	Classes	Training	Testing	Eval.Metric
Hallucination Assessment					
Object HalBench (Rohrbach et al., 2018)	2023	—	—	300	Accuracy
MMHal-Bench (Sun et al., 2023d)	2023	—	—	96	Hallucination Rate
LLaVA-Bench (In-the-Wild) (Liu et al., 2023b)	2023	—	—	60	GPT 4 based scoring
Image Captioning					
COCO Test 2017 (Lin et al., 2014)	2017	—	118,000	5,000	CIDEr / BLEU
Image Generation					
GenEval (Ghosh et al., 2023)	2023	6	—	533	Overall Score
Image Editing					
GEdit-Bench EN (Liu et al., 2025b)	2025	—	—	—	G_SC, G_PQ, G_O
GEdit-Bench CN (Liu et al., 2025b)	2025	—	—	—	G_SC, G_PQ, G_O

ingful step toward measuring this.

Efficiency-Quality Trade-off Metrics. Latency and parameter count are reported inconsistently across papers. MobileCLIP and MobileCLIP2 report latency on specific hardware (iPhone 12 Pro Max) alongside parameter counts, enabling meaningful efficiency comparisons. By contrast, generative VLM distillation papers typically report only parameter counts and LLM backbone size, omitting FLOPs and inference latency entirely. A standardized reporting protocol - parameters, FLOPs, and latency on a reference device - would substantially improve comparability across methods and make efficiency-quality trade-off curves tractable to construct.

D.4 Training Data and Evaluation Paradigm Evaluation.

Evolution of Training Data. As demonstrated in Table 7, VLM training data exhibits a growing structural separation between generic pre-training and task-aware post-training that reflects a fundamental shift in what knowledge distillation is actually transferring. The earliest datasets: SBU Caption, COCO Caption, Conceptual Captions established the basic image-text supervision paradigm, where scale was the primary lever, driving successive scaling from CC3M to LAION-400M to LAION-5B under the assumption that more diverse web-crawled pairs would produce richer student representations. However, web-crawled corpora carry systematic noise and a bias toward short factual descriptions that do not reflect the complex reasoning chains a capable teacher exhibits: raw scale could not substitute for curation quality. This motivates a decisive shift: modern distillation depends on task-specific, curated datasets. LLaVA-mix-665K provides structured instruction-following pairs; UltraFeedback and RLHF-V supply preference feedback that transfers the teacher's quality judgements rather than its token distributions. As the distillation objective evolved, the instruction dataset became the primary determinant of what behavioral capabilities get transferred. A parallel trend is the emergence of domain-specific distillation corpora. Remote sensing is the clearest example: general corpora are supplemented with SkySenseGPT and GeoChat-Instruct, reflecting that VLMs trained on internet imagery encode a systematic bias against satellite perspectives. Distilling from such teachers without domain-adapted supervision transmits that bias alongside the knowledge - a fundamental limitation of black-box sequence distillation that dataset specialization directly addresses.

Table 11: Summary of datasets used for VLM evaluation (Part V: remote sensing benchmarks). "—" denotes not applicable or not reported.

Dataset	Year	Classes	Training	Testing	Eval.Metric
RS Image Classification					
NWPU-RESISC45 (Cheng et al., 2017)	2017	45	—	31,500	Top-1 Acc
AID (Xia et al., 2017)	2017	30	—	10,000	Top-1 Acc
SIRI-WHU (Zhao et al., 2016)	2016	12	—	2,400	Top-1 Acc
WHU-RS19 (Xia et al., 2010; Dai and Yang, 2011)	—	19	—	950	Top-1 Acc
fMoW (Christie et al., 2018)	2018	63	—	16,948	Top-1 Acc
EuroSAT (RS) (Helber et al., 2019)	2019	10	—	27,000	Top-1 Acc
METERML (Zhu et al., 2022)	—	—	—	—	Top-1 Acc
RS Image-Text Retrieval					
UCM-Caption (Qu et al., 2016)	2011	—	—	2,100	R@1, R@5
RSICD (Lu et al., 2018)	2017	—	8,734	10,921	R@1, R@5
SkyScript (Wang et al., 2024d)	2024	—	—	—	Top-1 / R@1
RS Visual Question Answering					
RSVQA-HR (Lobry et al., 2020)	2020	—	—	—	Accuracy
RSVQA-LR (Lobry et al., 2020)	2020	—	—	—	Accuracy
RS Visual Grounding					
VG-DIOR (Zhan et al., 2023)	2023	—	—	—	Acc@IoU>0.5
RS General Knowledge					
LHRS-Bench (Muhtar et al., 2024)	2024	11	—	—	Accuracy

Evolution of Evaluation Paradigm. The evolution of evaluation benchmarks reveals what the field progressively discovered it was failing to measure. Early distillation was evaluated on ImageNet, CIFAR, and COCO - benchmarks inherited from computer vision and vision-language learning that primarily measured classification, retrieval, captioning, or detection performance. This framing was coherent for contrastive VLMs, where IN-val accuracy and Flickr R@1 meaningfully summarized capability preservation. However, it broke down for generative VLMs. A student matching its teacher's IN-val accuracy might still hallucinate objects, or produce fluent but factually incorrect responses. The shift was a transition toward holistic benchmarks that each target a distinct failure mode: MME disaggregates perception and cognition; MMMU introduces expert-level multidisciplinary reasoning that resists pattern matching; HallusionBench targets the entanglement between language priors and visual evidence that standard VQA metrics cannot expose. Each benchmark's emergence can be read as a post-hoc diagnosis: the community observed a failure mode in distilled models, constructed an instrument to measure it, then used that instrument to identify the behavioral fidelity. This redefinition of success has a direct implication for distillation methodology: a method that recovers VQAv2 accuracy while degrading POPE or MMMU scores has not successfully distilled the teacher: it has selectively compressed some capabilities while discarding others. The shift to multi-dimensional evaluation is therefore not a measurement convention but a claim about what distillation in the VLM setting is required to preserve.

E Knowledge Distillation for Vision-Language-Action Models

Knowledge distillation in Vision-Language-Action (VLA) models has become a general mechanism for improving data quality, representation richness, reasoning-action transfer, and deployment efficiency. Viewed from this perspective, recent VLA distillation methods can be organized by where knowledge is injected into the pipeline: the training distribution, the perceptual and action representation, the policy backbone, or the inference process

itself.

At the data level, distillation is increasingly used to replace or refine expensive human demonstrations. RLDG (Xu et al., 2024a) shows that trajectories generated by converged RL specialists can serve as stronger supervision than teleoperated data while preserving the generalization benefits of the VLA backbone. PLD (Xiao et al., 2025) further addresses the mismatch between offline training data and rollout-time failure states by training residual recovery policies and distilling their corrective behaviors back into the generalist. In a more explicitly data-centric formulation, FT-NCFM (Chen et al., 2025a) compresses large training corpora into compact, high-value coresets through fact tracing and adversarial synthesis, reducing training cost without substantial performance loss. Together, these works suggest that in VLA, distillation often acts first as a tool for curating better supervision.

A second line of work leverages distillation to enrich representations beyond what standard RGB-conditioned VLAs can capture. LatBot (Li et al., 2025d) learns transferable latent actions from large-scale manipulation videos and distills them into downstream VLAs, thereby injecting motion- and physics-aware priors before task-specific adaptation. GLaD (Guo et al., 2025) and FD-VLA (Zhao et al., 2026) extend the same principle to geometry and force, respectively, distilling privileged 3D or tactile information into latent tokens so that the deployed policy remains sensor-light at test time. VISTA (Chen et al., 2026b) addresses a related issue from the alignment side, first improving visual grounding in a surrogate task and then distilling that improved alignment into the instruction-following policy. In all cases, distillation serves as a bridge from auxiliary signals to a compact control policy that can operate under realistic sensing constraints.

Distillation is also increasingly used to manage the tension between high-level reasoning and low-level action execution. VITA-VLA (Dong et al., 2025) transfers action competence from a smaller action-specialized model into a large VLM through hidden-state alignment and selective fine-tuning. DualVLA (Fang et al., 2025) makes this trade-off explicit, arguing that naïve joint tuning can harm manipulation ability, and therefore uses dual-teacher supervision to preserve both reasoning and actuation performance. In the reverse direction, RPD (Jülg et al., 2025) distills a generalist VLA policy into a compact RL student, which can then continue improving through online interaction. These methods collectively frame distillation as a way to allocate competence across heterogeneous policies, rather than assuming a single model should learn reasoning and control equally well from scratch. Finally, a growing body of work treats distillation as a practical enabler of real-time deployment. Shallow-π (Jeon et al., 2026), MoLe-VLA (Zhang et al., 2025c), CEED-VLA (Song et al., 2025), and AC^2-VLA (Yu et al., 2026b) all use distilled supervision to preserve policy quality while reducing latency, either through backbone compression, dynamic layer execution, consistency-based decoding, or structured sparsification. Beyond tabletop manipulation, the same principle extends to broader embodied settings: OmniReason (Liu et al., 2025a) distills spatiotemporal reasoning for interpretable autonomous driving, while RDT2 (Liu et al., 2026) uses multi-stage distillation to support real-time, cross-embodiment generalization. In conclusion, the recent literature suggests that knowledge distillation in VLA should be understood less as a narrow compression recipe and more as a unifying strategy for transferring privileged data, sensing, structure, and policy competence into deployable embodied agents.

F Knowledge Distillation for Vision–Language Embedding Models

Recent advances in vision–language (VL) embedding models increasingly leverage knowledge distillation from large multimodal models (LVLMs or MLLMs) to transfer richer semantic and ranking signals into efficient embedding architectures. Instead of relying solely on contrastive training, these approaches exploit stronger teachers such as cross-encoder rerankers or multimodal LLMs to provide fine-grained supervision. For instance, Qwen3-VL-Embedding and Qwen3-VL-Reranker employ a two-stage retrieval framework in which a powerful cross-encoder reranker produces soft relevance scores for query–candidate pairs; these scores are then distilled into the embedding model via distribution alignment over positive and hard-negative samples, enabling the bi-encoder to approximate reranking-level semantic discrimination while maintaining efficient retrieval (Li et al., 2026). Similarly, UniME-V2 treats an MLLM as a semantic judge that assigns soft matching scores to query–candidate pairs drawn from a global candidate pool. The student embedding model is trained

to align its similarity matrix with these soft judgments, effectively distilling high-level multimodal reasoning into a lightweight embedding representation (Gu et al., 2025b).

Other work explores complementary distillation strategies to enhance the capability and generality of VL embeddings. xVLM2Vec introduces a self-knowledge distillation framework in which an LVLM-based embedding model serves as both teacher and student, enabling multilingual adaptation by aligning embeddings across translated image–text pairs while preserving the original English semantic structure (Musacchio et al., 2025). More broadly, recent studies explore using multimodal LLMs as teachers for universal embedding learning. By combining textual discriminative distillation with multimodal instruction tuning, these approaches transfer high-level semantic knowledge from large multimodal models into compact embedding representations that generalize across tasks such as retrieval, classification, and compositional reasoning (Gu et al., 2025a). Together, these works highlight a growing paradigm in which large multimodal models supervise lightweight embedding systems, enabling them to inherit richer cross-modal semantics and stronger ranking capabilities.

In general, distillation for VL embedding models aims to align the embedding space produced by a student model $f_S(\cdot)$ with that of a stronger teacher model $f_T(\cdot)$. A common objective is feature-level alignment, which encourages the student embeddings to match teacher embeddings:

$$\mathscr{L}_{\text{feat}} = \mathbb{E}_x \, \|f_S(x) - f_T(x)\|_2^2, \tag{9}$$

where x denotes an image–text input pair and f_T, f_S represent the teacher and student embedding functions respectively. When the teacher provides fine-grained ranking signals (e.g., from a reranker or MLLM judge), the student can also learn by matching the teacher's relevance distribution over a candidate set $\mathscr{D}$. Let $s_T(d_i|q)$ and $s_S(d_i|q)$ denote the teacher and student scores for document d_i given query q. The teacher distribution is defined using a temperature τ:

$$P_T(d_i|q) = \frac{\exp(s_T(d_i|q)/\tau)}{\sum_j \exp(s_T(d_j|q)/\tau)}, \tag{10}$$

and the student is trained to match this distribution via cross-entropy:

$$\mathscr{L}_{\text{rank}} = -\sum_i P_T(d_i|q) \log P_S(d_i|q). \tag{11}$$

These objectives allow lightweight embedding models to inherit semantic structure and ranking signals from stronger multimodal teachers, which has proven effective for improving retrieval-oriented VL representations.

G VLM Knowledge Distillation for Segmentation

Vision–language models (VLMs) trained on large-scale image–text corpora offer rich semantic knowledge that can be distilled into dense prediction models such as semantic segmentation networks. Models such as CLIP, trained on large-scale image–text corpora, encode rich open-vocabulary representations that can serve as powerful teachers for segmentation models. Instead of relying solely on pixel-level annotations, recent approaches leverage VLMs to provide supervision through feature distillation, pseudo-label generation, or language-guided alignment. These strategies allow segmentation systems to inherit the semantic knowledge learned from large-scale multimodal pretraining, enabling improved generalization and open-vocabulary recognition.

A common strategy is to extract dense supervision signals from VLM representations. For instance, MaskCLIP converts CLIP feature maps into pixel-wise pseudo-labels that can supervise segmentation networks without manual mask annotations (Zhou et al., 2022). Similarly, CLIP-based attention and similarity maps can provide localization cues for weakly supervised segmentation models (Lin et al., 2023). Recent work further demonstrates that dense segmentation masks can be extracted directly from VLM representations using plug-and-play pipelines, allowing segmentation models to be trained without dense annotations (Luo et al., 2023). In addition, large foundation models can generate high-quality pseudo-labels that serve as effective supervision signals in weakly supervised segmentation settings (Jiang and Yang, 2023). These approaches can be viewed as forms of knowledge distillation, where semantic information embedded in pretrained VLMs is transferred to task-specific segmentation networks.

Another line of research focuses on adapting frozen VLM representations for open-vocabulary semantic segmentation. Methods such as ZegCLIP (Zhou et al., 2023), Open-Vocabulary Semantic Segmentation with Frozen Vision-Language Models (Ma et al., 2022a), and Mask-Adapted CLIP

(Liang et al., 2023b) demonstrate that pretrained CLIP encoders can provide strong semantic supervision when integrated with segmentation architectures. Simple baselines further show that aligning pixel embeddings with textual category embeddings allows segmentation models to recognize unseen classes during inference (Xu et al., 2022). These ideas are extended to large-scale training setups that exploit image-level supervision to scale open-vocabulary segmentation models (Ghiasi et al., 2022). Unified frameworks such as FreeSeg also incorporate VLM supervision into more flexible segmentation pipelines supporting universal and open-vocabulary segmentation tasks (Qin et al., 2023).

Beyond feature transfer, language-guided and prompt-based approaches further exploit VLM knowledge for segmentation. Language-driven segmentation frameworks utilize textual descriptions to guide dense prediction models (Li et al., 2022a), while prompt-based methods combine text and image prompts to produce flexible segmentation outputs (Lüddecke and Ecker, 2022). Prompt-learning strategies have also been proposed to enhance open-world segmentation by learning category-level prompts that better align visual and textual representations (Zhu et al., 2023b). In addition, recent work demonstrates that VLM supervision can support segmentation without human-provided labels (Chen et al., 2023a) and even enable instance-level segmentation without manual mask annotations (VS et al., 2023). More explicitly, distillation-based approaches such as CLIPSelf show that vision transformers can distill knowledge from VLM representations to improve open-vocabulary dense prediction (Wu et al., 2024b). The distillation paradigm has also been extended beyond 2D images; for example, PartDistill transfers knowledge from vision–language models to perform 3D part segmentation (Umam et al., 2024). Overall, these studies highlight the potential of VLM knowledge distillation as a scalable and label-efficient strategy for building open-vocabulary segmentation systems.

H VLM Knowledge Distillation for Open-Vocabulary Object Detection

Knowledge distillation from pretrained vision-language models (VLMs), especially CLIP (Radford et al., 2021), has become one of the dominant paradigms for open-vocabulary object detection (OVD) (Zhang et al., 2024b; Wu et al., 2024a; Zhu and Chen, 2024; Li et al., 2025b). Early work established the basic transfer recipe: ViLD (Gu et al., 2021) distilled image- and text-level semantics from a pretrained VLM into a two-stage detector, while HierKD (Ma et al., 2022b) extended this idea to one-stage architectures through hierarchical global- and instance-level distillation. A central limitation of this paradigm, however, is the mismatch between image-level VLM supervision and region-level detection features. Much of the subsequent literature can be understood as attempts to narrow this gap. OV-OD (Bangalath et al., 2022) performs object-centric alignment and augments supervision with pseudo-labels; CLIP-Self (Wu et al., 2023d) aligns dense region features with crop-level CLIP representations through self-distillation; OADP (Wang et al., 2023a) and BARON (Wu et al., 2023c) enrich the transfer unit from isolated proposals to object-aware or bag-of-regions representations; and DeCLIP (Wang et al., 2025a) improves local discriminability and spatial consistency by decoupling content and context features.

Recent work has focused less on whether VLM knowledge should be transferred and more on how to make that transfer detector-friendly. DK-DETR (Li et al., 2023b) adapts distillation to DETR-style detectors with semantic and relational objectives, while HD-OVD (Fu et al., 2025) and CAKE (Ma et al., 2025a) emphasize multi-granularity or category-aware semantic transfer. Other methods refine the source of supervision itself: CCKT-Det (Zhang et al., 2025a) reduces reliance on extra supervision through cyclic contrastive transfer, and LBP (Li et al., 2024b) shows that background regions can also carry useful semantic cues instead of being treated purely as noise while MMC-Det (Xu et al., 2024c) distilled multi-modal contextual knowledge via a teacher fusion transformer trained with diverse masked language modeling, supervising concept-region similarity through attention scores. Related extensions have carried the same distillation perspective into adjacent settings, including simplified transfer schemes such as LP-OVOD (Pham et al., 2024), open-world detection (Ma et al., 2025b), and domain adaptation (Ke et al., 2024). These works suggest that the main challenge in distillation-based OVD is not the availability of VLM knowledge, but how to convert image-level, context-rich semantics into detector-compatible region representations that preserve lo-

cal detail, inter-object relations, and background cues.

I Structured Comparison of Distillation Methods

To provide a comprehensive overview of the VLM distillation landscape, this section systematically categorizes existing methods across the key dimensions of our proposed taxonomy. The following tables flatten specific branches of the taxonomy to facilitate direct comparison of methodological choices, including model architectures, distillation signals, teacher accessibility constraints, and evaluation paradigms.

Specifically, Table 12 details response-based methods focusing on logit distillation. Table 13 summarizes feature-based approaches that align intermediate hidden states. Table 14 outlines techniques operating under strict black-box teacher accessibility. Finally, Table 15 presents methods designed to bridge cross-modal structural mismatches.

Table 12: Structured Comparison of Logit-Distillation Methods

Method	Teacher	Student	Signal	Access	Data	Eval
LLaVA-KD (Cai et al., 2025)	Generative VLM	Smaller generative VLM	Logit + similarity distillation	White-box	1.2M multimodal instruction samples	VQA, MME, MMBench, POPE, MMMU
MASSV (Ganesan et al., 2025)	Target VLM	Draft VLM	Self-data + logit matching	White-box	Teacher-generated multimodal data	Speculative decoding benchmarks
Align-KD (Feng et al., 2024)	MobileVLM V2 7B	MobileVLM V2 1.7B	Logit + cross-modal alignment	White-box	ShareGPT4V, SQA, TextVQA	GQA, SQA, TextVQA, MME, MMBench
GenRecal (Lee et al., 2025a)	Large VLM	Small heterogeneous VLM	Logit + feature recalibration	White-box	Large multimodal corpora	AI2D, ChartQA, MathVista, MMMU
EM-KD (Feng et al., 2025b)	Efficient large VLM	Compact MLLM	Vision semantic + affinity distillation	White-box	Vision-language understanding data	Vision understanding and parsing benchmarks
PromptKD (Li et al., 2024c)	Large CLIP teacher	Prompt-based CLIP	KL-based logit distillation	White-box	Unlabeled domain images	Zero-shot classification, DG
KDPL (Mistretta et al., 2024)	CLIP/VLM teacher	Prompt learner	Prompt-level logit distillation	White-box	Unlabeled target-domain images	Domain generalization, classification
CLIP-KD (Yang et al., 2024)	Large CLIP teacher	Small CLIP student	Similarity/logit distillation	White-box	CC3M, LAION-style data	Zero-shot classification, retrieval
AMMKD (Li et al., 2025c)	Multiple CLIP teachers	Lightweight retrieval model	Multi-teacher KL distillation	White-box	Image-text retrieval datasets	Image-text retrieval
VL2Lite (Jang et al., 2025)	Large VLM	Lightweight compact model	Multimodal logit/feature distillation	White-box	Image classification datasets	Image classification
VLM-KD (Zhang et al., 2024d)	Off-the-shelf VLM	Long-tail classifier	Free-form text/logit distillation	White-box	Long-tail recognition datasets	Long-tail visual recognition
DHO (Kang et al., 2025a)	VLM teacher	Compact classifier	Teacher prediction distillation	White-box	Semi-supervised image datasets	ImageNet, fine-grained classification
MoPD (Chen et al., 2025b)	Hard prompt teacher	Soft prompt student	Prompt mixture distillation	White-box	Domain classification datasets	Prompt-based classification
Switch-KD (Sun et al., 2026)	Generative VLM (3B)	Generative VLM (0.5B)	Visual-switch logit distillation	White-box	Multimodal instruction data	Multimodal benchmarks
Prompt-stable KD (Malla et al., 2026)	CLIP-based teacher	Prompt-tuned student	Prompt stability + logit consistency	White-box	Unlabeled target-domain images	Zero-shot generalization
VLM-KD-IQA (Hou and Song, 2025)	CLIP-based IQA teacher	Compact IQA student	Modality-adaptive KD	White-box	IQA datasets	Image Quality Assessment
Hide-to-See (Yu et al., 2026a)	Strong reasoning VLM	Compact reasoning VLM	Prefix masking + logit distillation	White-box	Multimodal reasoning data	Multimodal reasoning benchmarks

Table 13: Structured Comparison of Hidden-State Distillation Methods

Method	Teacher	Student	Signal	Access	Data	Eval
Layerwised Multimodal KD (Wang et al., 2024b)	Large VLM (e.g. UNITER)	Small VLM	Layer-wise hidden alignment + modality losses	White-box	Multi-task VL corpora	VQA, NLVR, ITR, VE
KD-VLP (Liu et al., 2021)	Large VLM w/ object features	One-stream VLM	Object-guided MVM + phrase-region alignment	White-box	VG, COCO, SBU	VQA, NLVR, ITR
SF-CLIP (Sameni et al., 2024)	Foundation VL models	CLIP ViT-B/16	Masked KD with dense patch supervision	White-box	YFCC15M + CC12M	Zero-shot classification, retrieval, segmentation
TinyViT (Wu et al., 2022)	Large ViT	Tiny ViT	Distillation-token pretraining	White-box	ImageNet-1K	Image classification
DistillVLM (Fang et al., 2021)	Large VLM w/ detector	Small VLM	Attention + feature alignment	White-box	VL-7M corpus	COCO captioning, VQA
MiniVLM (Wang et al., 2020)	Large VLMs	Compact VLM	Pseudo-label distillation + efficient encoder	Black-box	COCO, VG, OpenImages	Captioning, VQA, ITR
VITA-VLA (Dong et al., 2025)	Action-expert VLA	VLM + action token	Action-space distillation	White-box	Driving/manipulation data	LIBERO robot benchmarks
Shallow-π (Jeon et al., 2026)	Full-depth VLA	Shallow VLA	Layer-wise feature KD + trajectory imitation	White-box	Simulated robotics data	Manipulation success rate
CEED-VLA (Song et al., 2025)	Full VLA	Early-exit VLA	Consistency KD + mixed-label supervision	White-box	Driving instruction data	Simulated and real robot tasks
CLIPSelf (Wu et al., 2023d)	CLIP ViT	Dense-feature CLIP	Crop-patch cosine alignment	White-box	COCO, VG	Open-vocab detection/segmentation
LP-OVOD (Pham et al., 2024)	CLIP (RN50)	Faster R-CNN + CLIP	Similarity distillation via pseudo labels	Black-box	COCO train split	Open-vocab detection
VLM-OOD (Li et al., 2023c)	Large CLIP	Small vision model	Hidden-state KD for OOD robustness	White-box	Multi-domain images	OOD image classification
Masters (Lee et al., 2025b)	InternVL3.5 38B	Smaller VLM (2B/4B)	Masked teacher KD + RL	White-box	Multimodal reasoning data	MMBench-style reasoning tasks
PracticalDG (Chen et al., 2024e)	CLIP	Small vision model	Perturbation distillation	White-box	PACS, Office-Home, etc.	Domain generalization
DILP (Kuang et al., 2023)	Large VLM	Prompt-enhanced VLM	Prompt-interaction hidden KD	White-box	VL pretraining corpora	Retrieval, VQA
Caption-Guided Replay (Lv and Li, 2026)	MLLM + T2I model	Continual-learning VLM	Multi-layer feature alignment	White-box	Sequential image streams	Continual learning benchmarks
LinMU (Wang and Jha, 2026)	Transformer VLM	Linear-attn VLM	Hidden/logit regression	White-box	CC, WebQA, etc.	VQA, long-context video QA
ProDistill (Luo et al., 2026)	Fine-tuned teacher	Merged student	Progressive layer-wise KD	White-box	Few-shot task data	Vision/NLP few-shot tasks
Drive-KD (Lian et al., 2026)	Multi-capability VLMs	Small driving VLM	Multi-teacher attention distillation	White-box	DriveBench logs	Autonomous driving tasks

Table 14: Structured Comparison of Black-box Teacher Accessibility Distillation Methods

Method	Teacher	Student	Signal	Access	Data	Eval
Visual Program Distillation (Hu et al., 2023)	Proprietary VLM APIs	Visual program executor	Program-trace distillation	Black-box	Teacher-generated reasoning programs	Visual reasoning, compositional VQA
ShareGPT4V (Chen et al., 2024c)	GPT-4V	Conversational VLMs	Instruction-response distillation	Black-box	GPT-4V-generated image conversations	MMBench, MMMU, MM-Vet, VQA
MiniGPT-4 (Zhu et al., 2023a)	BLIP-2 + ChatGPT supervision	Lightweight aligned VLM	Caption and conversational supervision	Black-box	Image-caption and dialogue data	Captioning, multimodal chat, VQA
ALLaVA (Chen et al., 2024a)	GPT-4V	Open-source LVLM	Synthetic instruction distillation	Black-box	Generated multimodal instruction data	Instruction following, VQA
PCoreSet (Kang et al., 2025b)	Proprietary VLM	Compact downstream VLM	Prediction-guided coreset distillation	Black-box	Unlabeled multimodal coreset	Visual understanding, adaptation
CLIP-Embed-KD (Nair, 2024)	CLIP embedding API	Lightweight vision encoder	Embedding-space distillation	Black-box	Image-text embedding datasets	Retrieval, zero-shot classification
RLDG (Xu et al., 2024a)	Large VL teacher	DG student model	Response-level distillation	Black-box	Multi-domain image datasets	Domain generalization
HierKD (Ma et al., 2022b)	Open-vocabulary detector	Compact OV detector	Hierarchical semantic distillation	Black-box	OV detection datasets	Open-vocabulary object detection
OV-OD (Bangalath et al., 2022)	CLIP-based OV teacher	OV detector	Cross-modal semantic distillation	Black-box	Detection + VL pretraining data	Open-vocabulary object detection
LVLM2P (Lee et al., 2025c)	Large LVLM	Prompt-based compact LVLM	Prompt distillation	Black-box	Teacher-generated instruction data	Multimodal instruction following
Online ICD (Kang et al., 2026)	Proprietary multimodal API	In-context learner	Online response distillation	Black-box	Low-resource multimodal tasks	Few-shot multimodal reasoning
Switch-KD (Sun et al., 2026)	Generative VLMs	Compact generative VLM	Visual-switch response distillation	Hybrid	Multimodal instruction tuning data	MMBench, MMMU, VQA, MME

Table 15: Structured Comparison of Cross-modal Structural Compatibility Distillation Methods

Method	Teacher Type	Student Type	Distillation Signal	Teacher Access	Training Data	Eval Task
A Sentence Speaks (Huang et al., 2023)	CLIP image-text model	Smaller vision model	Align image embedding to CLIP text embedding (cross-modal distance)	White-box	Multiple domain generalization datasets (PACS, VLCS, etc.)	Domain generalization (unseen-domain classification)
Enabling Multi-modal Gen. (Dai et al., 2022)	CLIP dual-stream VLM	BART (auto-regressive LLM)	Vision-language distillation (contrastive + text-text distance)	White-box	Conceptual Captions (3M image-text pairs)	Zero-shot VQA, zero-shot captioning
Distilling CLIP w/ Dual Guidance (Liu et al., 2024b)	CLIP image-text model	ResNet-50 visual encoder	Linguistic body-shape guidance + MSE visual feature distillation	White-box	Synthetic 3D human shape images + ReID images	Cloth-changing person re-ID
Unsupervised KD via Local Repr. (Baek et al., 2025)	Large VLM (e.g. CLIP-L/14)	Compact vision encoder	Global and local feature distillation (contrastive, patch alignment)	White-box	Unlabeled image datasets	Zero-shot classification, representation learning
TE-VLM (Liang and Shi, 2025)	ResNet-50 (vision model)	ResNet-34 (smaller vision model)	Logit distillation + Transfer Entropy reward	White-box	Unlabeled images (e.g. Food-101)	Zero-shot image classification (Food-101)
GLaD (Guo et al., 2025)	Frozen geometry-aware ViT (VGGT)	Vision-language action model (LLM with visual input)	Align LLM visual tokens with VGGT geometric features	White-box	Bridge (vision-language-action) dataset	Vision-language action (LIBERO robotics tasks)
FD-VLA (Zhao et al., 2026)	Force sensor signals (contact data)	Vision-language action model	Force token distillation via query-conditioned module	White-box	Contact-rich manipulation dataset (vision + force)	Contact-rich manipulation tasks
OV-OD (Bangalath et al., 2022)	CLIP image-text model	Open-vocabulary object detector (e.g. Faster R-CNN)	Object-centric CLIP embedding alignment (pseudo-labeling)	White-box	Detection datasets (COCO, LVIS)	Open-vocabulary object detection
VLV (Zhang et al., 2025d)	CLIP image-text model	Vision backbone (e.g. ResNet for detection/segmentation)	Vision-language latent alignment (dense matching)	White-box	Image-text pairs or detection data	(Open-vocab) object detection or segmentation
LVLM2P (Lee et al., 2025c)	Large VLM (vision-language LLM)	Reinforcement learning policy network	Teacher action supervision (instruction distillation)	Black-box	Environment trajectories from vision tasks	Vision-language reinforcement learning tasks
Prompt-stable KD (Malla et al., 2026)	CLIP-based vision-language model	Prompt-tuned classification model	Logit distillation + prompt stability regularization	White-box	Unlabeled image datasets (target domains)	Domain generalization, zero-shot classification
VLM-KD-IQA (Hou and Song, 2025)	Fine-tuned CLIP for IQA	Small IQA model (CNN)	Quality-prompt guided distillation (modality-adaptive)	White-box	Image quality assessment datasets	Image quality assessment
FLARE (Xie et al., 2026)	Pre-trained VLM (e.g. CLIP)	Driving model encoder (latent predictor)	Future latent prediction (self-supervised, no language)	White-box	Unlabeled driving trajectory data	Autonomous driving control
DAIT (He et al., 2026)	Large vision-language model	Smaller vision-language model	Cross-modal feature alignment (transformers)	White-box	Large-scale image-text corpus	VQA, captioning