EFFICIENT DENSE VISUAL REPRESENTATION LEARNING VIA MULTI-GRANULARITY SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual representations encode the semantic and visual content of real-world images in a feature space. They are widely used in semantic segmentation, cross-modal retrieval, and Multimodal Large Language Models (MLLMs). However, existing vision encoders, such as CLIP, are predominantly trained via image–text contrastive learning. This approach relies heavily on large-scale web-crawled alt-text as supervision, resulting in low data-efficiency during training. Moreover, the global instance-level contrastive objective limits the model's ability to capture fine-grained semantics and their corresponding relationships. In this paper, we propose TRIDENT, a data-efficient vision encoder pretrained with complementary forms of supervision: (i) instance-level image–text contrastive learning with a fine-tuned LLM as the text encoder to achieve cross-modal alignment and transfer its rich world knowledge to the vision encoder; (ii) region-level feature distillation to capture semantic consistency and spatial structure; and (iii) pixel-level reconstruction to preserve local visual details. Together, these objectives provide complementary global, regional, and pixel-level supervision, enabling richer dense visual representations and more data-efficient pretraining. Trained on only 1B samples, TRIDENT broadly outperforms eight competitive vision encoders trained on 4–40B samples across image–text retrieval, dense prediction, and downstream MLLM benchmarks. Compared with SigLIP2, TRIDENT improves I2T/T2I retrieval by 12.2/18.2 R@1 on Urban1K and 17.8/20.9 R@1 on ShareGPT4V, while achieving gains of 4.8 mIoU on ADE20K and 6.1 mIoU on PASCAL VOC for semantic segmentation. On MLLM downstream tasks, TRIDENT outperforms SigLIP2 on 10 of 11 benchmarks while matching it on MMMU, including gains of 3.4 points on VQAv2, 3.0 points on GQA, and 7.3 points on MMBench-EN. We will release the corresponding code and model checkpoints to facilitate further research in visual representation learning.

1 INTRODUCTION

Vision encoders serve as the foundation for image-text retrieval, dense prediction, and Multimodal Large Language Models (MLLMs). However, these applications require representations with complementary properties: retrieval relies on cross-modal semantic alignment, dense prediction demands spatially coherent features, and multimodal reasoning depends on fine-grained visual evidence (Radford et al., 2021; Zhai et al., 2023a). Consequently, a general-purpose vision encoder must integrate global semantic understanding with localized structural and appearance cues. Despite substantial progress enabled by large-scale pretraining, learning broadly transferable representations increasingly requires billions of training examples. This trend motivates a central question: *Can vision pretraining become more sample-efficient by extracting richer supervision from each image?*

Existing pretraining objectives offer complementary sources of supervision. Image–text learning methods, such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023a), align visual and textual representations using image-level correspondence. Self-distillation approaches, including DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2023), learn visual features by enforcing consistency and predictability in representation space. In contrast, masked image modeling methods such as MAE (He et al., 2021) supervise the reconstruction of masked visual content. Alongside advances in data scale, curation, and training recipes, these objectives have substantially improved representation quality. More recent approaches, such as SigLIP 2 (Tschannen et al., 2025), com-

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

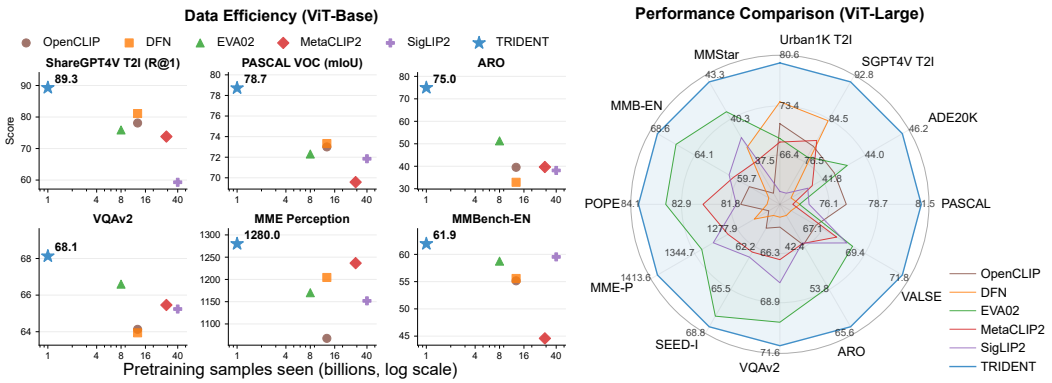

Figure 1: Data efficiency and performance comparison between TRIDENT and existing methods.

bine image–text learning with self-distillation and masked feature prediction. Collectively, these developments motivate a systematic investigation of how complementary supervision signals can be integrated to learn transferable representations under a limited sample budget.

Rather than solely scaling pretraining data, we improve data efficiency by extracting complementary supervision from each image–text pair: global cross-modal correspondence, regional semantic and spatial structure, and fine-grained visual appearance. We propose **TRIDENT**, which jointly optimizes instance-level image–text contrastive learning with a fine-tuned LLM text encoder (Huang et al., 2024), region-level feature distillation, and pixel-level masked reconstruction. As shown in Figure 1, TRIDENT trained on only **1B** samples outperforms eight competitive encoders pretrained on 4–40B samples across retrieval, dense prediction, compositional understanding, and downstream MLLM benchmarks. Compared with SigLIP2, it improves I2T/T2I R@1 by 12.2/18.2 on Urban1K and 17.8/20.9 on ShareGPT4V, mIoU by 4.8 on ADE20K and 6.1 on PASCAL VOC, and outperforms SigLIP2 on 10/11 MLLM benchmarks, with gains of 7.3 on MMBench-EN, 3.7 on MMStar, and 3.4 on VQAv2. The main contributions are summarized as follows:

- We identify **supervision granularity** and **per-sample information utilization** as key factors in data-efficient visual representation learning. By exploiting complementary supervision signals from each image–text pair at multiple granularities, our approach reduces the need to improve representation quality solely through larger pretraining corpora.
- We propose **TRIDENT**, an efficient dense visual representation learning framework that combines **instance-level** image–text contrastive learning, **region-level** EMA-teacher feature distillation, and **pixel-level** masked image reconstruction in a shared vision backbone.
- We evaluate TRIDENT on image–text retrieval, dense prediction, compositional understanding, and downstream MLLM benchmarks. Trained on only **1B** samples, TRIDENT outperforms eight competitive encoders pretrained on 4–40B samples.

2 Related Work

2.1 Visual Representation Learning

Large-scale visual representation learning has been extensively studied in both self-supervised learning and vision–language pretraining. In self-supervised learning, masked image modeling methods, such as MAE (He et al., 2021) and EVA02 (Fang et al., 2024b), learn visual representations through image reconstruction or feature prediction, whereas DINO (Caron et al., 2021) acquires semantic representations via self-distillation across augmented views. In vision–language pretraining, CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023a) learn from large-scale image–text pairs by aligning visual and textual representations. Subsequent work, including OpenCLIP (Ilharco et al., 2021), DFN (Fang et al., 2024a), and MetaCLIP2 (Chuang et al., 2026), has improved this paradigm through larger-scale training and enhanced data curation. More recently, AIMv2 (Fini et al., 2025) and GenLIP (Fang et al., 2026) have explored autoregressive and generative objectives for learning general-purpose vision encoders. Together, these methods demonstrate the effectiveness of diverse

supervision mechanisms for transferable visual representation learning, but largely emphasize scaling data and training rather than maximizing the information extracted from each sample. In contrast, TRIDENT improves data efficiency by jointly exploiting complementary supervision signals within every image–text pair.

2.2 Multi-Objective Visual Pretraining

Recent work has increasingly explored the integration of multiple supervision signals to enhance vision encoders. DINOv2 (Oquab et al., 2023) combines image-level self-distillation with masked patch-level feature prediction. SigLIP2 (Tschannen et al., 2025) augments image–text learning with captioning supervision, self-distillation, and masked feature prediction. FG-CLIP (Xie et al., 2025) enriches cross-modal learning through long captions, region–text alignment, and fine-grained hard negatives. MTV (Di et al., 2026) jointly optimizes vision–language alignment and visual self-supervision with dense spatial objectives, using expert-generated region-grounding and depth targets. Collectively, these methods demonstrate that complementary objectives can improve visual representation quality. In contrast, TRIDENT explicitly organizes supervision across instance, region, and pixel levels, thereby maximizing information utilization from each image–text pair and improving pretraining data efficiency.

3 Method

TRIDENT aims to improve the sample efficiency of dense visual representation learning through multi-granularity supervision of a shared vision encoder. As illustrated in Figure 2, given an image–text pair (x, t), we jointly train the encoder f_θ with objectives at multiple granularities. *(i) Instance-level contrastive learning* aligns global image representations with text representations from a fine-tuned LLM, providing rich cross-modal semantic supervision. *(ii) Region-level feature distillation* uses an Exponential Moving Average (EMA) teacher to provide semantic distillation across different views to enforce view-invariant semantics, and structural distillation within the same view to recover spatial structure. *(iii) Pixel-level reconstruction* uses a lightweight decoder to reconstruct the input image from visual features, encouraging the encoder to retain fine-grained visual information. Together, these complementary supervision signals exploit the information in each training sample across multiple levels of granularity, enabling richer visual representation learning with improved data efficiency.

3.1 Instance-Level Contrastive Learning

The instance-level contrastive objective establishes global cross-modal alignment between images and their associated captions. Inspired by recent works (Huang et al., 2024; Lee et al., 2025), we adapt Qwen3-8B (Yang et al., 2025) as a text encoder by enabling bidirectional attention and fine-tuning it with a caption-to-caption contrastive objective, after which it is kept frozen during vision-language pretraining. This adaptation provides LLM-derived language representations enriched with world knowledge to supervise visual pretraining through cross-modal alignment. Given an image–caption pair (x, t), the projected image and text representations are defined as:

$$v = g_v([f_\theta(x)]_{\mathrm{CLS}}), \qquad z = g_t(E_{\mathrm{LLM}}(t)), \tag{1}$$

where E_{LLM} denotes the adapted text encoder, g_v and g_t are the visual and text projection heads, and $[f_\theta(x)]_{\mathrm{CLS}}$ denotes the representation of the visual `[CLS]` token.

For each local mini-batch of B image–text pairs, we normalize the projected representations and gather them across all devices to form a global batch of N pairs. The image-to-text and text-to-image logits are computed as:

$$s_{ij}^{i2t} = \exp(\alpha)\, \bar{v}_i^\top \bar{z}_j + \beta, \qquad s_{ij}^{t2i} = \exp(\alpha)\, \bar{z}_i^\top \bar{v}_j + \beta, \tag{2}$$

where $\bar{v}$ and $\bar{z}$ denote the normalized image and text representations. The indices i and j refer to local anchors and global candidates. The learnable parameters α and β control the logit scale $\exp(\alpha)$ and bias, respectively.

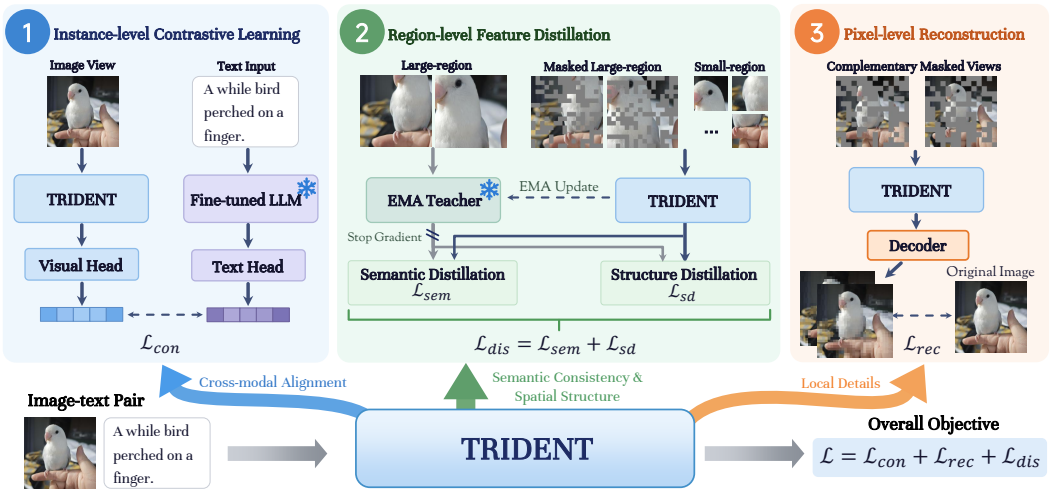

Figure 2: The architecture of the TRIDENT framework for efficient dense visual representation learning.

We use the sigmoid-based loss of SigLIP (Zhai et al., 2023b) as the instance-level contrastive objective:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B}\sum_{i=1}^{B}\sum_{j=1}^{N}\left[\log\sigma\left(y_{ij}^{i2t}s_{ij}^{i2t}\right) + \log\sigma\left(y_{ij}^{t2i}s_{ij}^{t2i}\right)\right], \tag{3}$$

where $y_{ij}^{i2t}, y_{ij}^{t2i} \in \{1, -1\}$ indicate whether the corresponding image and caption form a matched pair from image-to-text and text-to-image directions respectively. $\sigma(\cdot)$ is the sigmoid function.

3.2 Region-Level Feature Distillation

Instance-level contrastive learning supervises global correspondence but provides no explicit targets for individual spatial features. We therefore introduce region-level feature distillation using a teacher that shares the student vision encoder's architecture and is updated via an EMA of the student parameters. The teacher provides two complementary forms of supervision: *semantic distillation*, which promotes semantic consistency across views, and *structural distillation*, which encourages the preservation of spatial structure and contextual relationships among patch features within the same view.

3.2.1 Semantic Distillation

Semantic distillation promotes semantic consistency across region views of the same image. The teacher processes two unmasked views covering larger image regions (*large-region views*), while the student processes their masked counterparts and eight views covering smaller image regions (*small-region views*), yielding ten student views in total. Let $\mathbf{c}_T^{(a)}$ denote the teacher's `[CLS]` token representation for view $a \in \{1, 2\}$, and $\mathbf{c}_S^{(b)}$ the student's representation for view $b \in \{1, \ldots, 10\}$, where $b = 1, 2$ index the masked counterparts of the two teacher large-region views. Projection heads f_T and f_S map these representations into a shared semantic space. The resulting teacher and student distributions are defined as:

$$p_T^{(a)} = \text{softmax}\left(\frac{f_T(\mathbf{c}_T^{(a)}) - c_{\text{sem}}}{\tau_T}\right), \qquad p_S^{(b)} = \text{softmax}\left(\frac{f_S(\mathbf{c}_S^{(b)})}{\tau_S}\right), \tag{4}$$

where c_{sem} is the teacher centering vector, and τ_T and τ_S are temperature parameters.

Each teacher prediction supervises all student views except the one derived from the same large view, and the semantic distillation loss $\mathcal{L}_{\text{sem}}$ is defined as:

$$\mathcal{L}_{\text{sem}} = \frac{1}{18} \sum_{a=1}^{2} \sum_{\substack{b=1 \\ b \neq a}}^{10} H\left(p_T^{(a)}, p_S^{(b)}\right), \tag{5}$$

where $H(p, q) = -\sum_r p_r \log q_r$ denotes cross-entropy. The loss is averaged over the 18 valid teacher–student view pairs, with each teacher view supervising nine student views. Aligning predictions across region views encourages semantic consistency under variations in region size and visible context.

3.2.2 Structural Distillation

Structural distillation provides spatial supervision by training the student to predict teacher features at masked patch positions within the same region view. For each large-region view $k \in \{1, 2\}$, we independently sample a patch mask with a masking ratio of γ. The teacher processes the unmasked view, while the student processes its masked counterpart.

Let $\mathbf{h}_T^{(k,i)}$ denote the teacher feature at patch position i in view k, and $\hat{\mathbf{h}}_S^{(k,i)}$ the corresponding student feature computed from the masked view. The teacher and student patch projection heads, g_T and g_S, map these features into a shared feature space:

$$q_T^{(k,i)} = \text{softmax}\left(\frac{g_T(\mathbf{h}_T^{(k,i)}) - c_{\text{sd}}}{\tau_T}\right), \qquad q_S^{(k,i)} = \text{softmax}\left(\frac{g_S(\hat{\mathbf{h}}_S^{(k,i)})}{\tau_S}\right), \tag{6}$$

where c_{sd} is the centering vector applied to the teacher's patch-level logits. The structural distillation loss $\mathcal{L}_{\text{sd}}$ is evaluated only at the set of masked view–position pairs (k, i) using cross-entropy loss:

$$\mathcal{L}_{\text{sd}} = \frac{1}{|\mathcal{M}|} \sum_{(k,i) \in \mathcal{M}} H\left(q_T^{(k,i)}, q_S^{(k,i)}\right). \tag{7}$$

Predicting teacher distributions at masked positions encourages the student to infer missing patch representations from visible context, thereby modeling spatial dependencies among patch features. The final region-level distillation loss $\mathcal{L}_{\text{dis}}$ combines the semantic and structural losses:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{sd}}. \tag{8}$$

The two objectives provide complementary supervision: semantic distillation promotes cross-view semantic consistency, while structural distillation encourages the preservation of spatial structure within each view.

3.3 Pixel-Level Reconstruction

Region-level feature distillation supervises the student with teacher-derived feature targets, while pixel-level reconstruction provides targets directly from the input image. This complementary supervision encourages the vision model to preserve fine-grained appearance details that cross-modal alignment and feature distillation may not explicitly constrain.

Given an image x, we construct two complementary masked views $\tilde{x}^{(1)}$ and $\tilde{x}^{(2)}$, such that patches masked in one view are visible in the other. A shared lightweight pixel decoder D reconstructs the image from the spatial features of each view:

$$\hat{x}^{(1)} = D\left(f_\theta(\tilde{x}^{(1)})\right), \qquad \hat{x}^{(2)} = D\left(f_\theta(\tilde{x}^{(2)})\right). \tag{9}$$

For each view, the reconstruction loss is computed over its corresponding masked region using mean squared error:

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\left[\frac{1}{|\Omega^{(1)}|} \sum_{i \in \Omega^{(1)}} \left\|\hat{x}_i^{(1)} - x_i\right\|_2^2 + \frac{1}{|\Omega^{(2)}|} \sum_{i \in \Omega^{(2)}} \left\|\hat{x}_i^{(2)} - x_i\right\|_2^2\right], \tag{10}$$

Table 1: Visual representation evaluation on image-text retrieval, dense prediction, and compositional understanding tasks. OpenCLIP refers to the official checkpoints trained on DataComp-1B (Gadre et al., 2023). Best and second-best results within each block are highlighted in **bold** and underline, respectively. Entries marked with "–" indicate unavailable image-text evaluations because the corresponding models do not provide publicly available text encoders.

Model	Arch	Data	Image-Text Retrieval				Dense Pred.		Compos. Und.	
			Urban1K I2T	Urban1K T2I	SGPT4V I2T	SGPT4V T2I	ADE20K	PASCAL	VALSE	ARO
OpenCLIP (Ilharco et al., 2021)	B/16	12.8B	72.10	67.50	80.00	78.10	36.55	73.00	64.58	39.54
DFN (Fang et al., 2024a)	B/16	12.8B	**74.40**	68.00	85.00	81.10	36.72	73.33	65.07	32.88
EVA02 (Fang et al., 2024b)	B/16	8B	67.10	60.90	81.10	75.90	37.46	72.31	66.96	51.30
MetaCLIP2 (Chuang et al., 2026)	B/16	29B	67.20	61.00	81.30	73.80	35.41	69.58	66.06	39.67
SigLIP2 (Tschannen et al., 2025)	B/16	40B	42.40	43.70	58.70	59.30	36.53	71.85	68.23	38.14
TRIDENT	B/16	1B	72.10	**77.60**	**88.80**	**89.30**	**42.36**	**78.72**	**72.24**	**74.95**
AIMv2 (Fini et al., 2025)	L/14	12B	–	–	–	–	34.56	64.24	–	–
OpenCLIP (Ilharco et al., 2021)	L/14	12.8B	75.70	70.90	85.30	79.40	42.42	76.99	66.79	41.81
DFN (Fang et al., 2024a)	L/14	12.8B	75.70	74.00	87.00	84.20	40.22	74.46	65.45	33.90
EVA02 (Fang et al., 2024b)	L/14	4B	73.50	68.80	81.30	77.40	43.08	74.44	68.74	53.24
GenLIP (Fang et al., 2026)	L/16	8.0B	–	–	–	–	30.73	64.14	–	–
MetaCLIP2 (Chuang et al., 2026)	L/14	29B	**76.10**	68.30	83.30	80.50	41.30	74.07	67.89	41.43
CLIP (Radford et al., 2021)	L/14	12.8B	68.30	56.00	76.10	66.30	41.19	76.53	66.85	56.74
SigLIP2 (Tschannen et al., 2025)	L/16	40B	63.90	61.30	72.60	70.60	41.07	74.96	68.43	41.31
TRIDENT	L/16	1B	**76.10**	**79.50**	**90.40**	**91.50**	**45.88**	**81.08**	**71.40**	**63.78**

where $\Omega^{(1)}$ and $\Omega^{(2)}$ denote the masked regions in the two views, respectively.

Complementary masking allows each training sample to provide richer reconstruction supervision, thereby further improving the data-efficiency during training. By supervising directly with image content, this objective complements the preceding two objectives with signals for local appearance and fine-grained visual structure. The three supervision signals are jointly applied to the shared vision backbone without additional loss reweighting. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{dis}}. \tag{11}$$

4 Experiments

4.1 Implementation Details

We pretrain TRIDENT on a randomly sampled subset of 283M image–text pairs from Recap-DataComp-1B (Li et al., 2024b). The Qwen3-8B text encoder is first fine-tuned using LoRA on 50M Recap-DataComp caption pairs for one epoch, with a learning rate of 2×10^{-4} and a global batch size of 2048. TRIDENT is trained for 1B samples, corresponding to approximately 3.5 epochs over the 283M unique pairs. We train ViT-B/16 and ViT-L/16 models at an input resolution of 224×224, using global batch sizes of 5120 and 2560, respectively. We optimize both models with AdamW, using $(\beta_1, \beta_2) = (0.9, 0.95)$ and weight decay of 0.05. The learning rate linearly warms up to 6×10^{-4} over 800 steps, followed by cosine decay to 2×10^{-5}. Training uses bf16 mixed precision and gradient clipping with a maximum norm of 3.0. For structural distillation, we set the masking ratio to $\gamma = 0.4$. For masked reconstruction, we randomly mask 50% of image patches. The EMA teacher momentum follows a cosine schedule from 0.992 to 1.0.

4.2 Main Results

Cross-Modal Retrieval. Following LLM2CLIP (Huang et al., 2024), we evaluate zero-shot image–text retrieval on Urban1K (Zhang et al., 2024a) and ShareGPT4V-1k (Chen et al., 2023), reporting image-to-text (I2T) and text-to-image (T2I) Recall@1 (R@1). As shown in Table 1,

Table 2: Performance comparison on MLLM downstream benchmarks. MME-P and MME-C denote the raw perception and cognition scores, all remaining metrics are reported as accuracies (%).

Model	Arch	Data	General VQA			Perception & Hallu.			Comprehensive & Knowledge				
			VQAv2	GQA	SEED-I	MME-P	MME-C	POPE	MMB-EN	MMStar	MMT	MMMU	CV-Bench
OpenCLIP	B/16	12.8B	64.13	55.43	60.73	1067.62	307.68	79.75	55.15	34.88	44.45	39.44	62.02
DFN	B/16	12.8B	63.94	54.74	58.84	1204.40	260.71	79.38	55.58	35.21	42.05	39.67	58.26
EVA02	B/16	8B	66.60	58.01	63.62	1170.14	312.86	80.80	58.76	38.92	45.22	40.44	63.38
MetaCLIP2	B/16	29B	65.46	55.87	61.43	1236.56	269.29	81.07	44.59	34.96	43.24	37.22	60.20
SigLIP2	B/16	40B	65.24	56.99	60.93	1151.88	286.79	81.39	59.54	38.40	**45.47**	**41.11**	**64.48**
TRIDENT	B/16	1B	**68.13**	**58.93**	**66.32**	**1280.02**	**314.64**	**81.64**	**61.94**	**40.68**	45.00	40.56	63.61
AIMv2	L/14	12B	68.54	58.53	63.30	1287.71	284.29	83.26	63.40	39.55	**48.74**	38.78	62.43
OpenCLIP	L/14	12.8B	64.88	55.91	60.77	1228.49	338.57	81.59	58.51	35.35	44.93	40.33	62.70
DFN	L/14	12.8B	64.35	55.92	59.80	1251.14	336.79	80.99	56.53	38.42	44.55	41.33	61.60
EVA02	L/14	4B	69.94	59.06	67.49	1333.59	323.39	83.30	65.98	40.81	47.20	40.67	**67.48**
GenLIP	L/16	8.0B	70.32	60.27	**68.50**	1256.29	322.86	82.82	66.07	41.68	47.91	41.89	66.53
MetaCLIP2	L/14	29B	66.61	56.54	62.60	1292.50	320.36	82.45	60.14	37.44	46.59	39.78	63.91
CLIP	L/14	12.8B	64.95	56.45	57.53	1274.63	326.61	81.69	53.95	37.32	43.81	40.78	60.61
SigLIP2	L/16	40B	67.84	57.77	62.98	1315.53	310.36	81.69	60.57	39.08	45.54	**42.44**	61.75
TRIDENT	L/16	1B	**71.19**	**60.75**	68.30	**1403.12**	**380.71**	**83.92**	**67.87**	**42.81**	48.26	**42.44**	67.17

TRIDENT-L/16 achieves I2T/T2I R@1 scores of $76.10/79.50$ on Urban1K and $90.40/91.50$ on ShareGPT4V-1k. Despite training on only 1B samples, TRIDENT-L/16 surpasses SigLIP2-L/16, pretrained on 40B samples, by $12.2/18.2$ points on Urban1K and $17.8/20.9$ points on ShareGPT4V-1k. TRIDENT-B/16 exhibits a similar advantage under the same training budget. We attribute these gains to TRIDENT's multi-granularity objectives, which strengthen global image–text alignment while preserving region-level semantics and fine-grained visual details.

Dense Prediction. We evaluate spatial feature quality using the DINOv2 dense-probe protocol (Oquab et al., 2023) on ADE20K-150 (Zhou et al., 2017) and PASCAL VOC 2012 (Everingham et al., 2010), reporting mean intersection over union (mIoU). TRIDENT-L/16 achieves 45.88 mIoU on ADE20K and 81.08 mIoU on PASCAL VOC, exceeding SigLIP2-L/16 by 4.81 and 6.12 points, respectively. These consistent gains indicate that region-level distillation and pixel-level reconstruction yield spatially coherent representations that transfer effectively to dense prediction. These gains likely arise because region-level distillation explicitly preserves semantic consistency across spatial locations, while pixel-level reconstruction retains local appearance cues required for accurate dense prediction.

Compositional Understanding. We evaluate compositional understanding on VALSE (Parcalabescu et al., 2022) and ARO (Yuksekgonul et al., 2023). For VALSE, we report the unweighted mean pairwise-ranking accuracy across its 11 instruments on the validation split. For ARO, we report the unweighted mean over VG-Relation, VG-Attribution, COCO-Order, and Flickr30k-Order. TRIDENT-L/16 achieves scores of 71.40 on VALSE and 63.78 on ARO, indicating strong sensitivity to compositional changes in image–text correspondence. These results suggest that jointly modeling global cross-modal alignment and localized visual structure improves the encoder's ability to distinguish fine-grained relational and attribute-level differences.

MLLM Downstream Benchmarks. To assess transfer to multimodal understanding, we connect each pretrained vision encoder to Qwen3-8B through a MLP projector and follow the two-stage LLaVA training recipe (Liu et al., 2023): projector pretraining on LLaVA-558K, followed by multimodal instruction tuning on LLaVA-665K (Liu et al., 2024a). We evaluate the resulting MLLMs with LMMS-Eval (Li et al., 2024a; Zhang et al., 2024b) on VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), SEED-I (Li et al., 2023a), MME (Fu et al., 2026), POPE (Li et al., 2023b), MMB-EN (Liu et al., 2024b), MMStar (Chen et al., 2024), MMT-Bench (Ying et al., 2024), MMMU (Yue et al., 2024), and CV-Bench (Tong et al., 2024). SEED-I denotes the image-

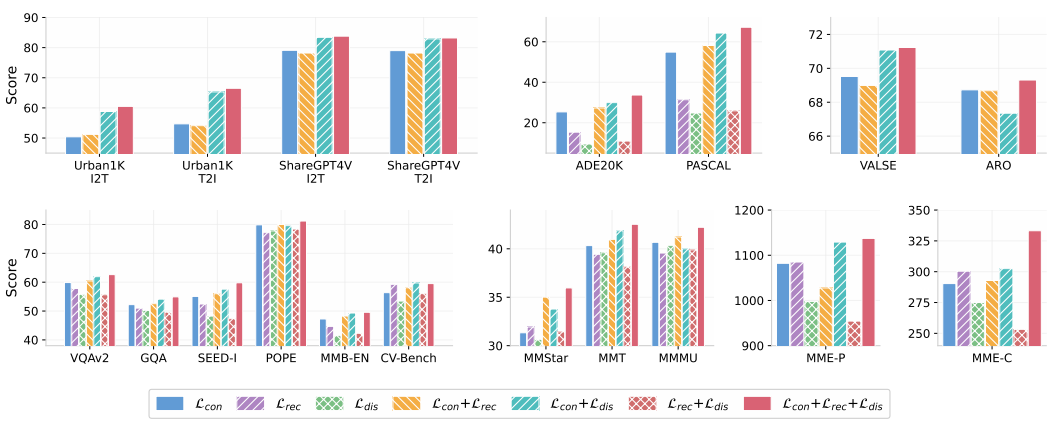

Figure 3: Ablation study on different losses. The top row reports visual representation evaluation, while the bottom row reports downstream MLLM evaluation. Urban1K, ShareGPT4V, VALSE, and ARO are not reported for configurations without $\mathcal{L}_{\text{con}}$, as these tasks require image–text alignment.

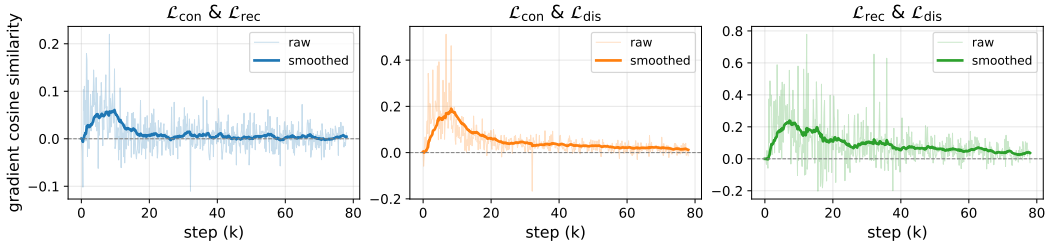

Figure 4: Gradient cosine similarity between the three training objectives during training. We report pairwise cosine similarities between loss gradients. All objective pairs are moderately aligned early in training and gradually approach zero, indicating increasingly near-orthogonal gradients without persistent conflict.

only subset of SEED-Bench, MMB-EN denotes static evaluation on the English development split of MMBench, and MME-P/C denote the perception and cognition scores of MME. As shown in Table 2, TRIDENT transfers effectively across both backbone scales. For ViT-B, it achieves the best performance on 8 of 11 metrics, including 68.13 on VQAv2, 58.93 on GQA, 66.32 on SEED-I, 1280.02/314.64 on MME-P/C, and 61.94 on MMB-EN. At the ViT-L scale, TRIDENT achieves 71.19 on VQAv2, 60.75 on GQA, 83.92 on POPE, 67.87 on MMB-EN, 42.81 on MMStar, and 1403.12/380.71 on MME-P/C. Compared with SigLIP2-L/16, TRIDENT-L/16 improves VQAv2, GQA, SEED-I, MMB-EN, and MMStar by 3.35, 2.98, 5.32, 7.30, and 3.73 points, respectively, while improving both MME-P and MME-C. These gains suggest that TRIDENT's multi-granularity supervision produces visual features that retain both global semantic information and localized visual evidence, facilitating their transfer to downstream multimodal reasoning.

4.3 Ablation and Analysis

4.3.1 Ablation on Different Losses

For all ablations, we pretrain ViT-B/16 for 100M sample presentations, varying only the enabled objectives while keeping all other settings identical to the main ViT-B/16 configuration. As shown in Figure 3, instance-level contrastive learning provides a strong cross-modal baseline. Adding region-level feature distillation improves Urban1K I2T/T2I R@1 from 50.40/54.70 to 58.80/65.30, and mIoU from 25.35 to 29.94 on ADE20K and from 54.89 to 64.30 on PASCAL VOC. Adding pixel-level reconstruction yields the strongest overall model, achieving the best results on all eight visual representation metrics and 10 of 11 MLLM metrics. Relative to contrastive learning with distillation, it further improves mIoU by 3.74 on ADE20K and 2.87 on PASCAL VOC, and improves MMStar and MMMU by 2.18 and 2.22 points. In contrast, reconstruction-only and distillation-only variants perform substantially worse on dense prediction. These results demonstrate the complementarity of cross-modal alignment, region-level distillation, and pixel-level reconstruction.

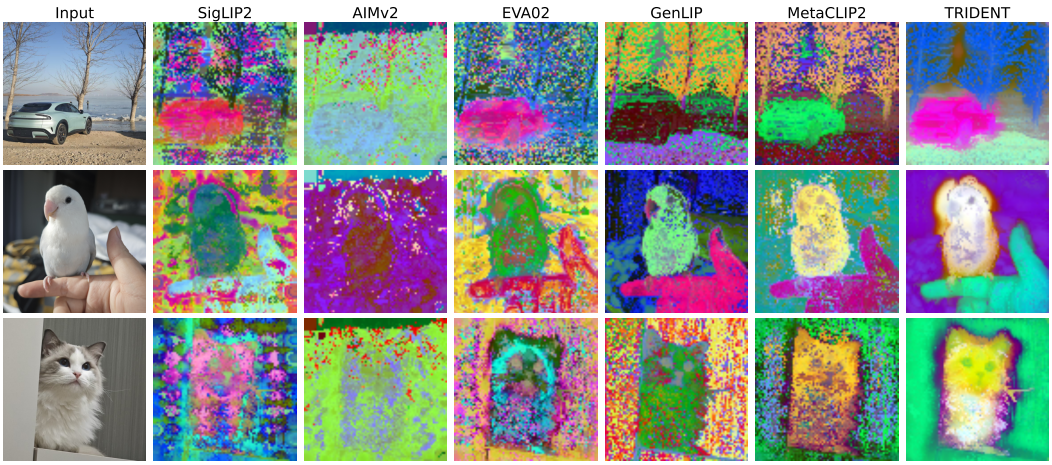

Figure 5: PCA visualization of dense features. Inputs are resized to 1280×1280 for models with a patch size of 16 and 1120×1120 for those with a patch size of 14, producing 80×80 feature maps in all cases. Similar colors denote semantically similar regions.

4.3.2 Gradient Direction Analysis

We analyze interactions among the three objectives by computing their gradients with respect to the shared vision trunk. Let $g_i(t) = \nabla_\theta \mathcal{L}_i(\theta_t)$ for $i \in \{\text{con}, \text{dis}, \text{rec}\}$ at step t. We measure pairwise alignment using cosine similarity. Figure 4 shows moderately positive similarities early in training, indicating partially aligned optimization directions. These similarities gradually approach zero, suggesting that the objectives become increasingly distinct without exhibiting persistent gradient conflict. Although near-orthogonal gradients alone do not establish feature-level complementarity, the ablation results confirm the empirical benefit of jointly optimizing the three objectives.

4.4 Qualitative Analysis

Figure 5 visualizes patch features from TRIDENT and five competing encoders by mapping the first three principal components to RGB channels. TRIDENT produces more spatially coherent feature maps: foreground objects form compact, consistently colored regions with sharper boundaries against the background. In contrast, competing encoders often exhibit noisy within-region variations and blurred object boundaries. This structure is consistent with TRIDENT's multi-granularity supervision: contrastive learning promotes semantic separation, region-level distillation encourages within-object consistency, and pixel reconstruction preserves boundary details. These observations align with TRIDENT's gains on dense prediction and downstream MLLM tasks.

5 Conclusion

In this paper, we introduce **TRIDENT**, a data-efficient framework for visual representation learning through multi-granularity supervision. TRIDENT makes more effective use of each image–text sample by combining three complementary learning signals: instance-level contrastive learning for global cross-modal alignment, region-level feature distillation for semantic consistency and spatial structure, and pixel-level reconstruction for fine-grained visual details. Jointly optimizing these objectives enables a shared vision encoder to learn richer representations under a limited pretraining sample budget. With only 1B training samples, TRIDENT achieves strong performance across image–text retrieval, dense prediction, compositional understanding, and downstream MLLM evaluation, outperforming competitive vision encoders trained with substantially larger sample budgets. These results highlight supervision granularity and per-sample information utilization as promising directions for improving the data efficiency of large-scale visual representation learning.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, assisting with translation and interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, designing or provide feedback on research methodology or experiments, cleaning and reformating dataset, supporting qualitative and thematic data analysis,and generating synthetic data sets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools to edit software code and improve the readability and clarity of the manuscript. All AI-assisted code and experimental suggestions were reviewed and verified by the authors, and all reported results and conclusions were independently checked against the actual experimental outcomes. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640. IEEE, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL https://arxiv.org/abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *NeurIPS*, 38:48009–48036, 2026.

Shangzhe Di, Zhonghua Zhai, and Weidi Xie. Revisiting multi-task visual representation learning. *arXiv preprint arXiv:2601.13886*, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *ICLR*, volume 2024, pp. 36221–36237, 2024a.

Yan Fang, Mengcheng Lan, Zilong Huang, Weixian Lei, Yunqing Zhao, Yujie Zhong, Yingchen Yu, Qi She, Yao Zhao, and Yunchao Wei. Let vit speak: Generative language-image pre-training. *arXiv preprint arXiv:2605.00809*, 2026.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, pp. 105171, 2024b.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *CVPR*, pp. 9641–9654. IEEE, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *NeurIPS*, 38, 2026.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *NeurIPS*, 36:27092–27112, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, Lili Qiu, and Liang Hu. Llm2clip: Powerful language model unlocks richer visual representation, 2024. URL https://arxiv.org/abs/2411.04997.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5143773. Version 0.1.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *ICLR*, volume 2025, pp. 79310–79333, 2025.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024a. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a. URL https://arxiv.org/abs/2307.16125.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023b. URL https://arxiv.org/abs/2305.10355.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024b. URL https://arxiv.org/abs/2307.06281.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL*, pp. 8253–8280, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *NeurIPS*, 37:87310–87356, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *ICML*, pp. 68777–68793, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. URL https://arxiv.org/abs/2404.16006.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11941–11952. IEEE, 2023a.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023b. URL https://arxiv.org/abs/2303.15343.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024a. URL https://arxiv.org/abs/2403.15378.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024b. URL https://arxiv.org/abs/2407.12772.

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