ReasonCLIP-58M: Visually Grounded Commonsense Reasoning Supervision for CLIP

Sicheng Zhang[1], Muzammal Naseer[1,2✉], Binzhu Xie[3], Naufal Suryanto[1], Shi Qiu[3], Jamal Bentahar[1], Naveed Akhtar[4], and Mubarak Shah[5]

[1] Khalifa University, UAE
[2] University of Western Australia, Australia
[3] The Chinese University of Hong Kong, China
[4] University of Melbourne, Australia
[5] University of Central Florida, USA

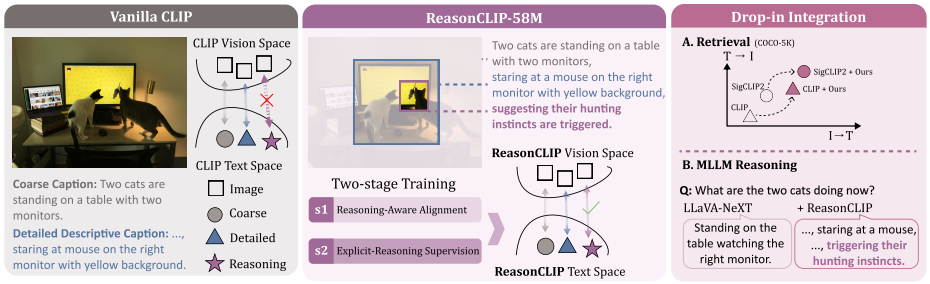

Fig. 1: ReasonCLIP-58M: From Description to Reasoning. ReasonCLIP-58M enables visually grounded reasoning in CLIP through two-stage continual pretraining with reasoning-aware and category-level supervision, improving zero-shot retrieval, structured reasoning, and MLLM performance without modifying the backbone.

Abstract. CLIP and its variants are widely adopted visual backbones in multimodal systems, but their pretraining remains dominated by descriptive image-text alignment. As downstream applications increasingly demand visually grounded commonsense inference and compositional reasoning, it remains unclear whether CLIP-style encoders can support such reasoning without architectural changes. To address this, we present ReasonCLIP-58M, a continual pretraining framework that integrates large-scale reasoning supervision into CLIP-style models through our two-stage strategy, which progressively integrates reasoning signals while preserving descriptive alignment, followed by category-structured reasoning supervision. To support this framework, we construct two complementary datasets and a benchmark: ReasonLite-42M, with open-form, visually verifiable reasoning captions; ReasonPro-16M, with category-specific reasoning supervision; and RCLIP-Bench for diagnostic evaluation of visually grounded reasoning. We train a family of ReasonCLIP

✉ Corresponding author. Contact: muhammadmuzammal.naseer@ku.ac.ae

that improves visually grounded commonsense and compositional reasoning while also enhancing zero-shot retrieval performance. As a drop-in visual encoder for multimodal large language models such as LLaVA-NeXT, ReasonCLIP delivers consistent gains without additional inference cost, demonstrating that structured reasoning supervision enhances the expressive capacity of CLIP-style visual representations. All datasets, models, and training code are available at `https://github.com/RISys-Lab/ReasonCLIP`.

1 Introduction

Contrastive Language-Image Pretraining (CLIP) [69] has emerged as a foundational vision-language model that aligns image and text representations through large-scale contrastive learning. Since its introduction, CLIP-style visual encoders [80, 86, 106] have become the default backbone for multimodal systems, including Multimodal Large Language Models (MLLMs) [1, 3, 49, 50]. Despite their widespread adoption, most CLIP models are pretrained on large-scale *descriptive* image-text datasets [36, 74], where supervision primarily encourages semantic correspondence rather than structured reasoning over visual content. While highly effective for retrieval, such training is not explicitly designed to support visually grounded reasoning [16, 37, 40, 87, 98, 110].

Advanced MLLMs increasingly require capabilities beyond descriptive alignment, including compositional reasoning [33, 39], commonsense inference [31, 103], and multi-step understanding [92, 96], when processing complex and diverse real-world scenarios.

However, as summarized in Tab. 1, recent CLIP advances largely focus on scaling descriptive data [26, 97, 106] or refining architectures [8, 47, 80, 107]. This emphasis creates a mismatch between pretraining objectives and reasoning-oriented downstream demands, questioning whether descriptive alignment alone is sufficient for visually grounded reasoning.

Motivated by this gap, we revisit CLIP and investigate whether its representation space can be extended through reasoning-oriented continual pretraining *without modifying its architecture*. As illustrated in Fig. 1, we propose ReasonCLIP-58M, a two-stage continual pretraining framework to train a family of CLIP-style visual encoders with over 58.9M visually grounded commonsense reasoning samples. This design progressively enhances reasoning capacity while explicitly preserving the model's original descriptive alignment.

In the first stage, we incrementally introduce reasoning signals to preserve the model's original descriptive alignment while enhancing its sensitivity to visually grounded inference. To support this, we construct ReasonLite-42M, which augments descriptive captions with visually verifiable reasoning statements. In the second stage, we strengthen supervision with structured category-level signals using ReasonPro-16M, a dataset spanning five visually grounded reasoning types: *Spatial/Geometric, Attribute/State, Creature/Action, Temporal/Phase, and Intuitive Physics*. This stage encourages the model to explicitly distinguish and organize diverse reasoning patterns within its visual representation space.

Table 1: Comparison of CLIP-style methods. Existing CLIP improvements typically follow two directions: descriptive data engineering (Rows. 1-6) and architectural refinement (Rows. 7-11). ReasonCLIP-58M training framework enables reasoning-oriented, non-intrusive continual pretraining while preserving backbone compatibility. Post. and Cont. denote post-training and continual pre-training, respectively. KG. denotes Knowledge Graph.

Model	Training	Dataset				Non-Intrusive			Supervision
		Name / Source	Construction	Scale	Reason	Text	Vision	Extra	
CLIP [69]	Pretrain	WIT	Web Crawl	0.4B	✗	-	-	-	Description
DataComp [26]	Pretrain	CommonPool	Web Crawl	1.4B	✗	✓	✓	✓	Description
MetaCLIP [15, 97]	Pretrain	MetaCLIP	Web Crawl	2.5B	✗	✓	✓	✓	Description
Eva-CLIP [23, 79]	Pretrain	LAION	Web Crawl	2.0B	✗	✓	✗	✗	Description
SigLIP [86, 106]	Pretrain	WebLI	Web Crawl	10.0B	✗	✓	✓	✓	Description
PE-Core [8]	Pretrain	MetaCLIP, PVD	MLLM Anno. (PVD)	5.4B	✗	✓	✗	✗	Description
Long-CLIP [107]	Post.	ShareGPT4V	N/A	1.0M	✗	✗	✓	✗	Long-Description
CLIP-MoE [111]	Post.	Recap-DataComp, ShareGPT4V	N/A	2.2M	✗	✗	✗	✗	Mixture-of-Experts
LLM2CLIP [32]	Cont.	DreamLIP, LAION, YMCC	N/A	60.0M	✗	✗	✓	✗	LLM-Augmented
READ-CLIP [46]	Post.	COCO	N/A	0.1M	✗	✓	✓	✗	Compositional Reasoning
DANCE [100]	Cont.	COCO, VG, SBU, CC3M, CC12M	KG. Anno.	14.1M	✗	✓	✓	✓	Knowledge Injection
ReasonCLIP	Cont.	CC12M	MLLM Anno.	58.9M	✓	✓	✓	✓	**Commonsense Reasoning**

Using this progressive training strategy, we train CLIP and SigLIP variants at multiple scales to validate cross-architecture generality, resulting in ReasonCLIP and ReasonSigLIP models, respectively. To systematically evaluate reasoning ability, we introduce RCLIP-Bench, a diagnostic benchmark that decomposes visually grounded reasoning into three hierarchical levels: visual grounding, evidence awareness, and structured reasoning.

Extensive experiments demonstrate that ReasonCLIP-58M consistently improves performance on visually grounded and compositional reasoning benchmarks across multiple backbones and model scales. Notably, reasoning-oriented continual pretraining also yields consistent gains in zero-shot retrieval tasks, suggesting that CLIP's representation space remains highly extensible. Furthermore, when integrated as a drop-in backbone into MLLM, such as LLaVA-NeXT [57], ReasonCLIP's encoder delivers stable performance gains without additional inference overhead. Our contributions are summarized as follows:

- **Explicit visually grounded commonsense reasoning supervision.** We introduce ReasonCLIP-58M, a continual pretraining framework and datasets that injects visually grounded commonsense reasoning into CLIP-style encoders while preserving descriptive alignment and architectural compatibility.
- **Two-stage reasoning-aware continual pretraining.** We design a structured two-stage training scheme that decouples reasoning enhancement from descriptive alignment preservation and enables organized reasoning patterns without degrading the original representation space.
- **Reasoning-enhanced CLIP/SigLIP variants**. We train ReasonCLIP and ReasonSigLIP models at multiple scales that consistently improve visually grounded and compositional reasoning as well as zero-shot retrieval.
- **Comprehensive evaluation and RCLIP-Bench.** We evaluate on diverse reasoning tasks and introduce RCLIP-Bench, which diagnoses visual grounding, evidence awareness, and structured visual reasoning.

2 Related Work

Supervision for CLIP. Most efforts to improve CLIP [69] strengthen *descriptive language supervision* for image-text alignment. A dominant line scales and filters large image-text corpora, including OpenCLIP [34], ALIGN [36], EVA-CLIP [80], SigLIP [86, 106], and MetaCLIP [15, 97]. Beyond data scaling, supervision is enhanced from both modalities: vision-side improvements leverage higher-quality video data or architectural refinements (e.g., Perception Encoder [8], EVA-CLIP [23, 79, 80], Vitamin [10]), while text-side methods reduce caption noise, enrich descriptions, or strengthen text encoders (e.g., [11,13,32,38,43,44,53,58,88,91,95,107,108,116]). Despite these advances, CLIP remains primarily optimized for descriptive alignment rather than reasoning-oriented representation learning. Recent works [6, 29, 54, 64, 72, 105, 112, 115], including CLIP-Event [52], TripletCLIP [68], and READ-CLIP [46], introduce structured supervision or auxiliary objectives to improve compositional reasoning, typically via fine-tuning, while methods such as DANCE [100] incorporate external knowledge to enhance commonsense capabilities. However, large-scale visually grounded reasoning signals during pretraining remain underexplored.
Benchmarks for CLIP. The evaluation of CLIP and its variants spans multiple levels. Fundamentally, zero-shot performance is measured on standard classification benchmarks [5, 18, 24, 28, 45, 71, 90, 94], as well as zero-shot image-text retrieval benchmarks [56, 101]. Recent evaluations further extend to long-form [11, 107], detailed [65], and multilingual [82] retrieval settings. For dense prediction and grounding tasks, CLIP features are commonly evaluated on semantic segmentation [21,118], depth estimation [35,75], and referring expression comprehension [60,102]. To assess higher-level reasoning capabilities, vision-and-language association is evaluated using WinoGAViL [7], while compositional reasoning is evaluated using benchmarks that test attribute, object, and relational substitution [19, 30, 41, 59, 83, 104]. However, most existing CLIP benchmarks assess descriptive alignment or compositional correctness without disentangling distinct failure modes in visually grounded reasoning, motivating the introduction of RCLIP-Bench for fine-grained diagnostic evaluation.
More detailed related works are discussed in Appendix E.

3 ReasonCLIP-58M and RCLIP-Bench

Existing reasoning datasets are primarily designed for MLLMs and involve complex reasoning, making them unsuitable for direct training of CLIP-style encoders. To bridge this gap, we introduce a novel two-stage training framework that fosters progressive, visually grounded reasoning. To facilitate this, we constructed a specialized suite of datasets (ReasonLite-42M and ReasonPro-16M) that supports the distinct learning objectives of each stage and a new benchmark (RCLIP-Bench) that explicitly evaluates three hierarchical reasoning levels. The overall design and representative examples are illustrated in Fig. 2 with additional implementation details in Appendix A.

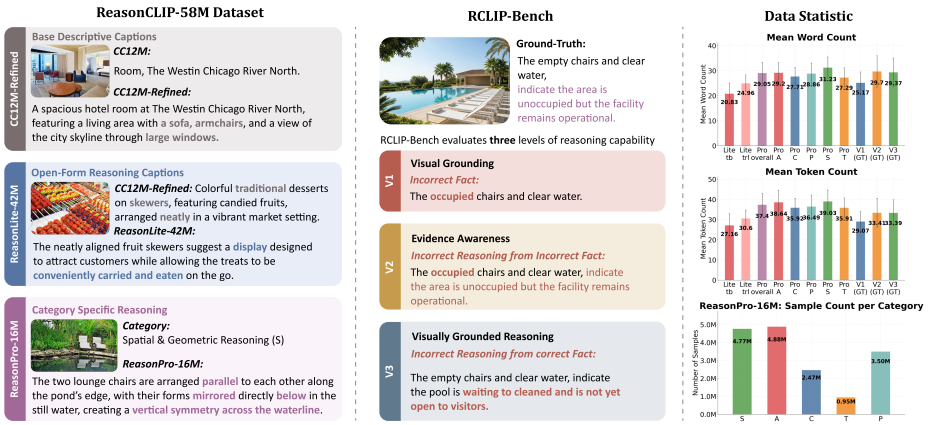

Fig. 2: Overview of ReasonCLIP-58M & RCLIP-Bench. *ReasonCLIP-58M* provides visually grounded reasoning supervision for CLIP-style models. Built from a refined version of *CC12M* [9], it comprises *ReasonLite-42M* for open-form reasoning and *ReasonPro-16M* for category-specific reasoning. We also introduce *RCLIP-Bench*, a dedicated benchmark for three-level evaluation of visually grounded reasoning. We report the token and word distributions for each dataset.

3.1 ReasonLite-42M Dataset: Open-Form Reasoning

ReasonLite is an *open-form reasoning* dataset containing diverse and verifiable reasoning statements grounded in visual evidence. These statements extend beyond factual descriptions and complement descriptive supervision with inference patterns that encourage reasoning-aware visual representations. As shown in Fig. 2, our reasoning statements leverage visually observable evidence (*"neatly aligned"*) in the image to infer higher-level design intentions and functional advantages (*"conveniently carried and eaten"*). To ensure data quality, we exclude samples that exhibit false causality, over-extension, hallucination, or complex multi-step reasoning beyond what is visually supported. More details are shown in Appendix A.1.

Dataset Construction. We use CC12M [9] as the image source and retain 10.4M valid images. Due to the relatively low quality of the original CC12M captions, we employ Qwen2.5-VL-72B [4] to regenerate three high-quality descriptive texts per image, ensuring factual consistency and descriptive completeness. We denote these regenerated captions as ***Text Base*** (T_b) and refer to the resulting dataset as *CC12M-Refined*. This step yields 31.2M relabeled image-text pairs. From CC12M-Refined, 4.7M images are used to construct ReasonLite, while the remaining 5.7M are reserved for ReasonPro.

We then input each image and its corresponding T_b into the annotation model and instruct it to extract key visual evidence grounded in T_b, producing an evidence set E. Within the same generation request, the model partitions E into three semantically coherent clusters $\{E_1, E_2, E_3\}$. For each cluster E_i, the model performs commonsense reasoning based on these cues and generates a corre-

Table 2: Five categories of perception-grounded reasoning types in ReasonPro-16M.

ID	Category	Definition
S	**Spatial / Geometric**	Reasoning about spatial relations (position, direction, distance, occlusion, containment) and how object placement or geometry affects visibility, reachability, or motion.
A	**Attribute / State**	Reasoning about intrinsic properties or visible conditions of objects, including appearance, texture, deformation, openness, or on/off states, based on visual cues.
C	**Creature / Action**	Reasoning about human or animal posture, gesture, and interaction with objects to infer current or imminent behavior.
T	**Temporal / Phase**	Reasoning about the temporal stage of an event (past, ongoing, or upcoming) based on motion continuity, trajectory, or dynamic context.
P	**Intuitive Physics**	Reasoning about intuitive physical relationships (stability, support, contact, forces) inferred directly from visual cues, without relying on semantic common sense.

sponding caption ***Text Reason Lite*** (T_{rl}). The T_{rl} captions express visually verifiable commonsense inferences derived from the grouped evidence, and the prompts explicitly constrain all reasoning statements to remain grounded in the image and consistent with T_b. In total, this process yields 42M image-caption pairs in **ReasonLite-42M**. Additional construction details, including statistics, prompt design, and computational cost, are provided in Appendices A.2 and A.3.

3.2 ReasonPro-16M Dataset: Category-Specific Reasoning

ReasonPro introduces category-specific commonsense reasoning captions to provide more structured and controllable supervision than the open-form generation in ReasonLite. While ReasonLite encourages diverse reasoning patterns, ReasonPro explicitly targets predefined visual reasoning types, enabling clearer supervision signals and fine-grained per-category analysis aligned for CLIP.

Reasoning Categories. In Table 2, we define five visually grounded reasoning patterns compatible with CLIP-style encoders: *Spatial/Geometric (S), Attribute/State (A), Creature/Action (C), Temporal/Phase (T), and Intuitive Physics (P)*. For each category, captions are constructed from category-specific descriptive cues such that the inferred commonsense reasoning remains explicitly grounded in the corresponding visual attribute, relation, action, temporal stage, or physical condition. Consistent with ReasonLite, all reasoning in ReasonPro is grounded in factual descriptions and observable visual evidence. Detailed definitions and examples for each category are provided in Appendix A.4.

Dataset Construction. All annotations in this stage are conducted with Qwen3-VL-32B [81], which provides stronger category discrimination and more structured outputs than the model used for ReasonLite. Starting from the 5.7M images reserved from CC12M-Refined, we first perform a multi-label category annotation stage. Given the predefined reasoning categories and their definitions, the annotation model assigns a label set $Y \subseteq \{S, A, C, T, P\}$ to each image, indicating the visually supported reasoning types. To ensure that each image supports multiple reasoning patterns and yields richer supervision, we retain only images with $|Y| \geq 3$, yielding 5.5M images for ReasonPro construction.

For each retained image, we randomly select three reasoning categories $Y' \subseteq Y$. We then feed the image, its corresponding T_b, the selected categories Y', and

the category definitions into the annotation model within a single generation request. The model generates one category-specific reasoning caption for each selected category, producing three ***Text Reason Pro*** (T_{rp}) samples per image. In total, ReasonPro contains 5.5M images and 16.6M image-caption pairs spanning up to five reasoning categories. Further details are provided in Appendix A.4.

3.3 RCLIP-Bench: Visually Grounded Reasoning Benchmark

RCLIP-Bench is a benchmark for fine-grained diagnostic evaluation of visually grounded commonsense reasoning. Although existing benchmarks effectively measure image-text alignment using descriptive captions [56, 78, 101], they are insufficient for assessing visually grounded reasoning. RCLIP-Bench evaluates three levels of reasoning capability: *Visual Grounding* (V1), *Evidence Awareness* (V2), and *Visually Grounded Reasoning* (V3). This hierarchical design enables precise diagnosis of where a model's reasoning fails.

Benchmark Construction. RCLIP-Bench is constructed from high-quality image–caption pairs in the **DOCCI** dataset [65], which provides detailed human-authored descriptions serving as reliable factual references. For each image, the original DOCCI caption is treated as the positive instance. We then generate three types of hard negatives corresponding to V1-V3. The benchmark follows the same five reasoning categories defined in ReasonPro-16M (S/A/C/T/P). As illustrated in Fig. 2, RCLIP-Bench adopts a three-tier hierarchy of hard negatives, each targeting a distinct failure mode:

- **Incorrect Facts (V1).** These captions contain factually inconsistent descriptions and directly test visual grounding ability. Within V1, errors are divided into five sub-types: *subject/object form*, *noun substitution*, *relational misuse*, *attribute confusion*, and *sentence structure alteration*. Failure at this level indicates insufficient perceptual accuracy.
- **Incorrect Reasoning from Incorrect Facts (V2).** These captions begin with an incorrect factual premise and derive flawed reasoning from it. It evaluates evidence awareness: whether the model can detect incorrect premises rather than being misled by a coherent but visually unsupported narrative.
- **Incorrect Reasoning from Correct Facts (V3).** These captions describe the scene factually correctly but draw an incorrect reasoning. This level tests whether a model can distinguish valid commonsense reasoning from invalid conclusions based on the same visual facts.

Evaluation is conducted in a contrastive retrieval setting, where the model must rank the correct caption above its corresponding hard negatives.

3.4 Quality Control

We adopt a two-stage quality-control process to ensure the reliability of the dataset. Prior work suggests that large-scale training is relatively robust to moderate noise [36]. Therefore, our focus is on preventing systematic errors and generation bias rather than eliminating all minor imperfections.

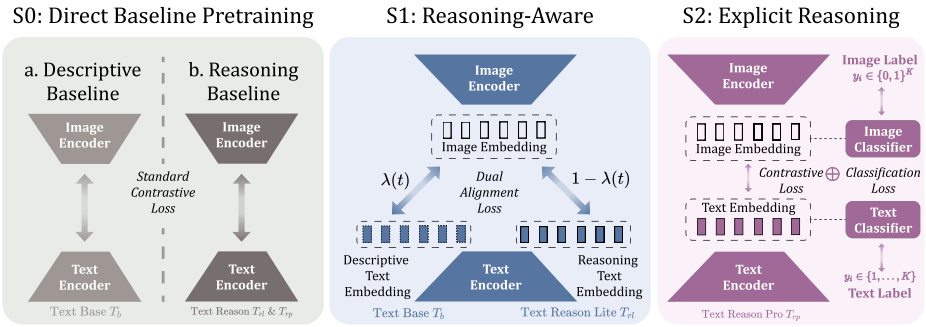

Fig. 3: Overview of the ReasonCLIP training framework. ReasonCLIP uses a stage-wise continual pretraining strategy that progressively integrates reasoning capacity into the visual encoder while preserving its foundational semantic alignment, enabling seamless integration as a drop-in backbone in MLLMs.

Manual Validation. Before large-scale generation, we conduct a pilot validation round with 500 samples per stage. Each sample is independently reviewed by five graduate students. Prompts are iteratively refined, and large-scale generation proceeds only after the pass rate exceeds 99.5%.

Automatic Filtering. For ReasonLite, we apply rule-based filtering to remove overly long, short, malformed, or degenerate generations, with a removal rate below 0.01‰. In ReasonPro, 3.97% (0.20M) of samples are filtered during category annotation due to inconsistent or low-confidence labels, and additional structural and grounding constraints are enforced during caption generation. After large-scale generation, we randomly inspect 500 samples again, achieving a pass rate above 99.0% across all stages. For RCLIP-Bench, quality control ensures clear separation and difficulty stratification across the three negative tiers (V1/V2/V3). Negative captions are generated with GPT-5 [76] under tier-specific prompts, followed by filtering to remove degenerate outputs, near-duplicates, and tier violations, resulting in a removal rate of approximately 13%. Additional quality control details are provided in Appendix B.

4 ReasonCLIP Training Framework

We propose a two-stage continual pretraining framework that integrates visually grounded reasoning into CLIP-style visual encoders while preserving descriptive alignment. As shown in Fig. 3, training consists of three stages: baseline pretraining (Stage 0), reasoning-aware alignment (Stage 1), and explicit category-level reasoning supervision (Stage 2). We further demonstrate drop-in integration of the resulting ReasonCLIP encoder into downstream multimodal systems (Stage 3), enabling stable adaptation without additional inference cost.

4.1 Stage 0: Baseline Continual Pretraining

Stage 0 establishes two baseline pretraining setups for comparison: purely descriptive alignment (*S0-Des.*) and naive reasoning supervision (*S0-Rea.*).

In the *S0-Des.* model, only descriptive captions T_b from CC12M-Refined are used for training. In the *S0-Rea.* model, all reasoning captions (including T_{rl} from ReasonLite and T_{rp} from ReasonPro) are directly mixed as training targets, without distinguishing reasoning types or stages. Both settings optimize the same standard image-text alignment objective. Formally, given an image-text pair (x, t), the model is optimized with a backbone-specific alignment loss:

$$\mathcal{L}^{(0)} = \mathcal{L}_{\text{align}}(x, t), \tag{1}$$

where $\mathcal{L}_{\text{align}}$ denotes the backbone-specific image–text alignment objective. For example, CLIP-style models adopt the symmetric InfoNCE loss, while SigLIP models use a sigmoid-based binary cross-entropy loss.

4.2 Stage 1: Reasoning-Aware Alignment

Stage 1 introduces visually grounded commonsense reasoning awareness into the CLIP-style model while preserving its original descriptive alignment.

We continue pretraining on the ReasonLite-42M dataset using both descriptive captions T_b and reasoning captions T_{rl} as supervision signals. This dual-supervision objective maintains the original alignment structure while gradually incorporating reasoning signals, encouraging the model to become sensitive to reasoning cues without disrupting its learned visual semantics. To prevent excessive parameter drift from the original backbone, we introduce an ℓ_2 regularization term toward the initial weights.

Formally, for a given image x with descriptive caption T_b and reasoning caption T_{rl}, the Stage 1 objective is defined as

$$\mathcal{L}^{(1)} = \lambda(t)\,\mathcal{L}_{\text{align}}(x, T_b) + \big(1 - \lambda(t)\big)\,\mathcal{L}_{\text{align}}(x, T_{rl}) + \beta\|\theta - \theta_0\|_2^2, \tag{2}$$

where $\mathcal{L}_{\text{align}}$ denotes the backbone-specific image–text alignment loss, $\lambda(t)$ is a time-dependent weight at training step t, θ and θ_0 represent the current and original model parameters, respectively, and β controls the regularization.

To gradually shift emphasis toward reasoning supervision, $\lambda(t)$ is defined as

$$\lambda(t) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \text{clip}\left(\frac{t - t_1}{t_2 - t_1},\ 0,\ 1\right). \tag{3}$$

In practice, $\lambda(t)$ decreases over time, progressively increasing the influence of reasoning captions while retaining descriptive alignment as the dominant signal in early training. This scheduling enables the smooth integration of reasoning signals within a unified representation space

4.3 Stage 2: Explicit Reasoning Supervision

Stage 2 builds upon the reasoning awareness established in Stage 1 and shifts the objective toward explicit category-level reasoning supervision.

To this end, Stage 2 training is conducted on the ReasonPro-16M dataset, where descriptive captions T_b are no longer used as supervision. The primary objective is to align images with category-specific reasoning captions T_{rp}. Since alignment alone does not explicitly encourage discriminability across different reasoning categories, we introduce a category-discrimination branch that imposes category-level supervision, thereby enhancing the separability of distinct reasoning patterns while maintaining a unified feature space [114].

Formally, for a given image x with corresponding reasoning caption T_{rp}, let $Y \subseteq \{S, A, C, T, P\}$ denote the multi-label reasoning categories associated with image x, and let c denote the single reasoning category of caption T_{rp}. The overall optimization objective in Stage 2 is

$$\mathcal{L}^{(2)} = \mathcal{L}_{\text{align}}(x, T_{rp}) + \gamma\big(\mathcal{L}_{\text{img-cls}}(x, Y) + \mathcal{L}_{\text{txt-cls}}(T_{rp}, c)\big), \tag{4}$$

where $\mathcal{L}_{\text{img-cls}}$ and $\mathcal{L}_{\text{txt-cls}}$ denote the image-side multi-label and text-side single-label classification losses, respectively; γ balances alignment and category-level supervision. These losses are defined as

$$\mathcal{L}_{\text{img-cls}} = \text{BCE}\big(g_v(x), \mathbf{y}\big), \quad \mathcal{L}_{\text{txt-cls}} = \text{CE}\big(g_t(T_{rp}), c\big), \tag{5}$$

with $\mathbf{y} \in \{0,1\}^5$ the multi-hot reasoning label vector of image x, and $c \in \{1, \ldots, 5\}$ the reasoning category of caption T_{rp}. The functions $g_v(\cdot)$ and $g_t(\cdot)$ are lightweight MLP-based classification heads attached to the image and text encoders, respectively, and are used only during training (discarded at inference).

4.4 Stage 3: Drop-in Integration

Stage 3 demonstrates the integration of ReasonCLIP into downstream multimodal systems. We apply the proposed continual pretraining framework to a CLIP-style backbone and replace the original visual encoder with the resulting ReasonCLIP encoder. The target system then proceeds with its standard alignment or fine-tuning procedure. Formally, let a multimodal system be denoted as $\mathcal{M}$ with visual encoder ϕ_{Baseline}. The integration can be expressed as

$$\mathcal{M}(x, \cdot) = \mathcal{F}\big(\phi_{\text{Baseline}}(x), \cdot\big) \longrightarrow \mathcal{F}\big(\phi_{\text{ReasonCLIP}}(x), \cdot\big), \tag{6}$$

where $\mathcal{F}$ denotes the remaining components of the multimodal system. This replacement enables downstream systems to leverage reasoning-aware visual representations without altering their existing pipelines.

5 Experiments

5.1 Implementation Details

Training Setting. Data generation and training are conducted on NVIDIA A100 64G GPUs, requiring around 3.8k and 3.5k GPU hours, respectively. Following Sec. 4, each stage is trained for one epoch on its corresponding dataset

Table 3: Zero-Shot Text-Image Retrieval Results. *All 31K images are used.

Model	Res.	Data	COCO-5K [56] I→T R@1	COCO-5K [56] I→T R@5	COCO-5K [56] T→I R@1	COCO-5K [56] T→I R@5	Flickr-30K [101] I→T R@1	Flickr-30K [101] I→T R@5	Flickr-30K [101] I→T R@10	Flickr-30K [101] T→I R@1	Flickr-30K [101] T→I R@5	Flickr-30K [101] T→I R@10	Urban-1K [107] I→T R@1	Urban-1K [107] I→T R@5	Urban-1K [107] T→I R@1	Urban-1K [107] T→I R@5	RCLIP-V3-5K (**Ours**) I→T R@1	RCLIP-V3-5K (**Ours**) I→T R@5	RCLIP-V3-5K (**Ours**) I→T R@10	RCLIP-V3-5K (**Ours**) T→I R@1	RCLIP-V3-5K (**Ours**) T→I R@5	RCLIP-V3-5K (**Ours**) T→I R@10
Scale - ViT Base (86M)																						
OpenCLIP [34]	224/32	0.4B	52.3	76.3	34.2	60.0	39.7	63.8	72.8	24.0	44.0	53.3	55.8	78.1	55.4	77.9	54.8	77.8	84.7	27.8	47.2	54.5
MetaCLIP [97]	224/32	0.4B	51.8	76.4	35.9	61.8	39.3	63.2	72.4	25.7	46.7	56.2	57.2	79.7	52.6	73.7	56.5	80.6	87.6	30.1	50.2	57.8
DataComp [26]	224/32	1.4B	53.4	77.5	37.2	62.3	39.0	62.7	71.9	24.6	44.9	54.1	64.4	85.9	59.9	88.3	**60.8**	**82.6**	**88.8**	32.2	52.0	59.3
CLIP [69]	224/32	0.4B	50.0	75.0	30.4	56.0	40.7	64.7	73.7	21.8	41.6	51.1	61.0	84.8	46.8	72.8	51.0	75.3	84.1	26.7	47.2	55.5
+ **Stage 1**	224/32	+42M	**56.2**	**78.8**	**37.9**	**64.1**	**42.5**	**68.0**	**77.2**	**29.1**	**51.1**	**60.7**	**70.4**	**91.6**	**68.6**	**90.2**	56.0	79.3	87.0	30.4	51.5	59.4
+ **Stage 2**	224/32	+16M	52.3	77.2	37.0	62.6	41.9	66.4	75.7	27.5	49.0	58.4	59.2	83.0	60.4	83.9	55.3	80.5	88.0	**33.8**	**56.0**	**64.0**
Scale - ViT Large (307M)																						
OpenCLIP [34]	224/14	2.0B	63.4	84.0	46.5	71.1	55.0	79.0	86.4	39.4	62.2	70.9	72.0	90.7	68.0	89.0	67.1	87.3	92.8	37.3	56.8	63.0
EVA-CLIP-02 [79]	224/14	2.0B	63.7	84.3	**47.5**	71.2	58.2	81.0	87.7	**42.2**	**64.7**	**73.0**	73.4	91.0	70.0	87.0	67.6	87.6	92.8	38.3	58.0	64.1
MetaCLIP [97]	224/14	2.5B	64.4	85.0	47.1	71.4	57.1	80.4	87.6	40.9	63.8	72.3	74.4	89.5	69.9	85.9	**69.1**	88.9	93.6	38.9	58.3	64.4
Long-CLIP [107]	224/14	0.4B	63.3	84.7	47.1	**71.9**	54.5	78.8	86.4	41.5	64.4	72.9	**82.8**	**96.6**	**86.1**	**96.4**	59.7	83.8	90.6	37.7	58.0	64.7
CLIP [69]	224/14	0.4B	56.3	79.5	36.6	61.2	48.8	73.0	81.1	28.2	49.6	59.0	68.5	88.9	55.9	80.4	54.0	77.9	86.1	31.7	52.3	59.5
+ **Stage 1**	224/14	+42M	**64.5**	**85.4**	46.7	71.6	**59.9**	**82.4**	**88.8**	40.9	63.7	72.4	80.5	96.2	79.7	94.2	66.3	86.2	92.1	38.0	58.4	65.1
+ **Stage 2**	224/14	+16M	59.9	82.5	46.2	70.8	55.6	78.8	85.8	39.0	61.8	70.5	74.3	92.9	75.5	91.5	67.1	**89.1**	**94.1**	**41.9**	**63.1**	**69.3**
ViTamin [10]	336/16	1.0B	64.3	85.0	47.3	71.0	56.1	79.6	86.6	39.9	62.3	70.8	80.6	93.7	76.0	93.2	**69.7**	88.3	92.8	39.8	58.5	64.1
EVA-CLIP-02 [79]	336/14	2.0B	64.2	85.2	47.9	71.7	59.7	82.1	88.7	43.4	66.0	74.1	77.1	91.9	70.7	87.8	67.9	88.0	93.0	39.1	58.7	64.6
SigLIP2 [106]	384/16	10.0B	**70.9**	**89.6**	**55.5**	**78.1**	**68.6**	**88.3**	**93.1**	**51.2**	**73.1**	**80.2**	66.9	86.4	61.9	80.3	66.5	82.3	86.9	41.8	59.5	65.4
CLIP [69]	336/14	0.4B	58.0	80.9	37.0	61.6	51.4	75.5	83.3	31.6	53.5	62.7	73.0	90.3	57.0	79.5	55.9	79.5	86.6	33.2	53.6	60.7
+ **Stage 1**	336/14	+42M	65.1	85.7	46.9	71.7	61.0	83.0	89.5	41.9	64.8	73.2	**83.0**	**96.3**	**81.9**	**94.5**	67.2	86.7	92.7	38.5	58.7	65.3
+ **Stage 2**	336/14	+16M	61.3	83.0	47.1	71.3	57.2	79.8	86.8	40.8	63.4	72.0	75.5	91.8	78.2	93.5	68.0	**89.9**	**95.0**	**42.8**	**63.4**	**69.8**
Scale - ViT So400m (428M)																						
SigLIP [106]	384/14	10.0B	72.6	90.1	54.3	76.8	69.9	89.3	93.9	51.3	73.0	80.2	74.5	91.9	73.4	88.3	70.5	86.3	**90.2**	43.4	61.6	66.8
+ **Stage 1**	384/14	+42M	73.6	**91.2**	56.5	79.2	69.3	88.8	93.8	52.3	73.8	80.9	77.5	**93.1**	**82.3**	**94.5**	68.3	84.4	87.7	45.5	63.7	69.3
+ **Stage 2**	384/14	+16M	**73.7**	90.8	57.3	79.9	73.3	90.0	94.3	53.9	75.3	82.3	**78.6**	92.3	80.6	93.8	**72.4**	**86.4**	89.5	**49.5**	**67.8**	**73.0**
SigLIP2 [86]	384/14	10.0B	71.2	89.7	55.7	78.5	69.9	89.3	93.9	51.3	73.0	80.2	64.1	84.7	65.0	84.1	64.9	78.2	80.7	38.5	54.8	60.0
+ **Stage 1**	384/14	+42M	72.8	89.8	58.2	80.4	69.6	89.0	93.6	53.6	74.8	81.7	75.0	92.8	75.5	91.0	64.5	77.7	81.0	40.5	56.7	61.8
+ **Stage 2**	384/14	+16M	73.3	89.9	**59.0**	**81.1**	**73.8**	**91.2**	**94.9**	**54.8**	**76.2**	**83.1**	73.9	91.7	75.5	92.3	68.4	80.5	82.7	43.8	60.6	65.2
Scale - ViT Giant Opt (1B)																						
SigLIP2 [86]	384/16	10.0B	72.4	**90.6**	56.5	78.6	69.0	89.0	93.7	53.0	74.4	81.3	58.6	78.3	58.2	78.2	65.2	79.5	**83.1**	40.0	57.0	62.5
+ **Stage 1**	384/16	+42M	72.0	89.5	59.3	81.3	71.4	90.1	94.5	55.5	76.6	83.3	**72.2**	**92.9**	**77.9**	**92.3**	65.1	78.4	81.1	43.6	61.1	66.6
+ **Stage 2**	384/16	+16M	**73.7**	89.3	**60.1**	**81.7**	**75.3**	**91.6**	**95.3**	**56.7**	**77.5**	**84.0**	71.0	90.4	75.5	90.0	**67.4**	**79.7**	81.9	**47.2**	**64.8**	**70.2**

with an effective batch size of 24,576 or 32,768, depending on the model scale. We train six CLIP [69] and SigLIP [86, 106] variants at different scales. More detailed training configurations for each stage are provided in Appendix C.

We integrate both CLIP-L/14-336 and ReasonCLIP-L/14-336 into the LLaVA NeXT [57] framework with Qwen3-1.7B [81] as the language backbone, keeping the vision tower frozen during training. Following the standard training pipeline, we use 558K samples for pretraining and 779K samples for fine-tuning.

Baselines. We adopt a tiered comparison strategy. First, we compare against the original backbones (CLIP and SigLIP) to measure performance gains from reasoning-oriented continual pretraining. Second, we include data-centric methods such as MetaCLIP [97] and DataComp [26]. We further compare with approaches involving limited structural modifications (e.g., Long-CLIP [107] and Eva-CLIP-02 [23]), as well as compositional reasoning methods such as READ-CLIP [46]. As a non-intrusive framework, to ensure fair comparison, we exclude large-scale architectural or LLM-augmented approaches such as EVA-CLIP-18B [80], PE-Core [8], and LLM2CLIP [32].

Evaluation. We re-evaluate all models under a unified protocol based on CLIP-bench [14]. Please refer to each result section and Appendix D.1 for details.

5.2 Results

Observation 1. *ReasonCLIP consistently improves model performance on both descriptive retrieval and reasoning retrieval tasks across backbones and*

Table 4: Commonsense reasoning results. †Human evaluation is conducted on 1,000 samples. In this table, we report Stage 1 Model results for our models. Full results for all training stages are shown in Appendix D.3.

Model	Res.	Data	WinoGAViL [7] Candidates			RCLIP-Bench **(Ours)** V1: Visual Grounding						V2: Evidence Awareness						V3: Visual Reasoning					
			5-6	10-12	Avg.	Att.	Nou.	Rel.	Sen.	Sub.	Avg.	S.	A.	C.	T.	P.	Avg.	S.	A.	C.	T.	P.	Avg.
Scale - ViT Base (86M)																							
OpenCLIP [34]	224/32	0.4B	53.2	42.6	50.6	21.7	26.6	**16.7**	**19.2**	29.1	22.7	17.6	18.5	18.3	28.2	20.8	20.7	18.5	26.3	22.0	18.8	23.7	21.8
MetaCLIP [97]	224/32	0.4B	**58.3**	**51.0**	**56.5**	20.0	27.5	16.1	18.5	29.9	22.4	12.6	15.3	19.8	26.9	17.3	18.4	16.7	21.7	**26.9**	**28.9**	19.5	22.7
DataComp [26]	224/32	1.4B	55.6	45.8	53.3	21.2	28.3	16.3	18.9	30.8	23.1	14.2	16.1	24.3	**34.9**	16.9	21.3	19.4	22.9	22.3	23.9	22.8	22.3
CLIP [69]	224/32	0.4B	53.5	41.5	50.6	22.4	28.9	16.2	17.0	27.2	22.3	12.9	15.4	11.3	26.5	15.2	16.2	**19.6**	23.7	21.6	22.5	**25.8**	22.6
ReasonCLIP	224/32	+**42M**	57.8	49.6	55.8	**22.5**	**31.4**	14.2	18.5	**31.4**	**23.6**	**18.2**	**19.3**	**25.0**	31.5	**21.3**	**23.1**	18.5	**31.6**	21.5	23.5	23.6	**23.8**
Scale - ViT Large (307M)																							
OpenCLIP [34]	224/14	2.0B	53.8	44.0	51.5	22.4	32.3	15.8	19.3	34.4	24.8	14.2	15.0	**25.5**	29.3	17.5	20.3	19.9	18.6	19.9	27.7	19.0	21.0
Eva-CLIP-02 [23]	224/14	2.0B	56.5	50.5	55.0	22.4	33.7	13.3	18.8	35.4	24.7	11.3	16.9	19.6	30.8	14.7	18.7	21.8	22.5	20.0	25.4	19.0	21.7
MetaCLIP [97]	224/14	2.5B	59.6	53.6	58.2	23.5	33.3	15.6	19.7	35.1	25.4	13.3	17.6	18.7	24.0	16.3	18.0	19.3	22.5	**25.5**	30.1	23.8	24.3
Long-CLIP [107]	224/14	0.4B	**63.7**	56.5	**62.0**	24.0	34.3	**16.4**	**20.4**	34.2	25.9	**21.0**	**28.0**	19.6	**35.6**	**22.2**	**25.3**	**25.4**	**36.5**	18.9	**31.8**	25.8	**27.7**
CLIP [69]	224/14	0.4B	53.3	41.1	50.4	20.4	31.1	13.9	17.5	29.8	22.5	8.2	13.0	12.5	32.1	20.8	17.3	23.4	25.8	24.1	21.2	**29.3**	24.8
ReasonCLIP	224/14	+**42M**	62.5	**56.6**	61.1	**25.8**	**36.7**	15.0	19.3	**36.0**	**26.6**	11.7	19.2	15.7	34.3	15.2	19.2	19.6	28.3	18.0	26.7	28.8	24.3
ViTamin [10]	336/16	1.0B	54.4	46.8	53.7	21.3	33.3	**15.4**	**20.1**	34.7	25.1	14.6	15.5	23.1	31.5	16.2	20.2	20.2	20.2	16.1	24.3	21.3	20.4
Eva-CLIP-02 [79]	336/14	2.0B	56.8	50.5	55.3	22.3	34.1	13.5	18.8	36.0	24.9	11.1	17.7	18.6	29.7	14.5	18.3	20.6	22.4	20.8	25.5	19.2	21.7
SigLIP2 [86]	384/16	10.0B	**63.2**	**58.4**	**62.1**	16.3	22.6	14.2	12.6	20.2	17.2	**15.9**	18.7	**24.8**	33.1	20.4	**22.6**	19.4	21.5	21.2	21.9	16.9	20.2
CLIP [69]	336/14	0.4B	53.0	41.4	50.2	21.3	31.7	13.7	17.6	30.3	22.9	8.2	13.3	13.5	32.1	**21.3**	17.7	**24.2**	26.0	**25.4**	21.2	28.3	25.0
ReasonCLIP	336/14	+**42M**	62.3	54.6	60.2	**26.4**	**37.5**	**15.4**	19.7	**36.8**	**27.2**	11.7	**19.8**	15.6	**34.3**	14.9	19.3	20.3	**29.7**	19.5	**28.8**	**28.6**	**25.4**
Scale - ViT So400m (428M)																							
SigLIP [106]	384/14	10.0B	63.0	57.7	61.7	24.0	33.3	14.3	21.2	29.3	24.4	13.0	17.3	28.0	30.3	22.2	22.2	20.6	25.6	17.7	**30.5**	14.5	21.8
ReasonSigLIP	384/14	+**42M**	58.7	50.9	56.8	**32.1**	**38.5**	**17.2**	**25.1**	**41.1**	**30.8**	16.1	29.1	28.8	24.9	**30.1**	25.8	**26.2**	**34.3**	21.0	23.9	24.5	**26.0**
SigLIP2 [86]	384/14	10.0B	62.1	56.8	60.8	15.7	19.9	14.3	12.1	17.6	15.9	17.6	19.1	**30.7**	**35.7**	18.3	24.3	20.5	23.6	**23.1**	25.0	16.2	21.7
ReasonSigLIP2	384/14	+**42M**	**65.2**	**60.1**	**64.0**	18.1	20.9	15.0	12.5	17.8	16.9	**17.2**	**27.5**	30.5	27.9	28.0	**26.2**	20.4	**34.3**	18.3	23.1	**26.1**	24.4
Scale - ViT Giant Opt (1B)																							
SigLIP2 [86]	384/16	10.0B	**64.4**	**61.5**	**63.7**	15.0	20.4	**15.2**	11.5	15.6	15.6	**16.1**	19.3	**32.1**	**28.7**	21.2	23.5	**20.2**	23.5	**20.5**	**29.8**	16.9	22.2
ReasonSigLIP2	384/16	+**42M**	62.1	60.1	61.6	**18.1**	**20.9**	15.0	**12.5**	**17.8**	**16.9**	**16.1**	**28.4**	27.8	24.8	**30.0**	**25.4**	20.1	**30.4**	15.9	25.6	**25.8**	**23.6**
Unbounded Scale																							
GPT-5.2	-	-	-	-	-	92.7	83.6	78.2	85.5	94.5	86.5	92.7	83.6	89.1	87.3	80.0	86.5	89.1	80.0	90.9	87.3	94.5	88.4
Human†	-	-	-	-	-	**96.5**	**95.1**	**95.3**	**97.1**	**94.7**	**95.7**	**93.0**	**89.7**	**93.9**	**88.7**	**90.0**	**91.1**	**92.2**	**85.8**	**94.0**	**90.1**	**96.2**	**91.7**

model scales. In Table 3, we report image-text retrieval results on COCO [56], Flickr [101], Urban1K [107], and the proposed RCLIP-Bench, covering standard retrieval, long-caption retrieval, and commonsense reasoning caption retrieval settings. ReasonCLIP and ReasonSigLIP both outperform their respective baselines in descriptive and reasoning retrieval, even including strong large-scale variants. Notably, despite being trained on a comparatively smaller reasoning corpus, our models surpass approaches that rely on substantially larger pretraining datasets or architectural modifications (e.g., MetaCLIP, Eva-CLIP-02, Long-CLIP). These results suggest that reasoning-oriented continual pretraining is effective for enhancing visually grounded representation quality.

Observation 2. *RCLIP-Bench provides a hierarchical evaluation of visually grounded reasoning.* As shown in Table 4, while WinoGAViL is informative for vision-language associations, high scores on this benchmark do not reliably reflect truly visual reasoning ability. RCLIP-Bench addresses this by decomposing evaluation into three hierarchical levels, revealing that no single model consistently dominates across all levels, and that strong retrieval performance does not guarantee strong reasoning. Notably, our ReasonCLIP consistently improves over its base CLIP backbone across all three levels, demonstrating that reasoning-oriented continual pretraining effectively enhances visually grounded reasoning ability beyond what WinoGAViL alone can capture.

Observation 3. *ReasonCLIP improves compositional reasoning across backbones without task-specific fine-tuning.* Table 5 compares compositional reasoning results across multiple benchmarks. Across different backbone architec-

Table 5: Compositional reasoning results. We report Stage 2 results for our models. Full results for all training stages are shown in Appendix D.4.

Model	ViT	Res.	Data	WhatsUp [41]	VALSE [67]	CREPE [59]	SugarCREPE [30]	SugarCrepe++ [19]		Avg.
								ITT	TOT	
MetaCLIP [97]	B/32	224	0.4B	46.5	69.1	18.2	76.9	58.1	50.6	53.2
CLIP [69]	B/32	224	0.4B	41.2	67.5	23.9	73.1	60.0	46.7	52.1
ReasonCLIP	B/32	224	**+58M**	51.3 ↑10.1	72.5 ↑5.0	24.0 ↑0.1	75.6 ↑2.5	61.5 ↑1.5	61.5 ↑14.8	57.7 ↑5.6
– Compositional Reasoning Methods: Fine-tuned on MS-COCO, 100K Samples										
Triplet-CLIP [68]	B/32	224	-	41.6	64.2	15.0	82.7	61.7	57.4	53.8
GNM-CLIP [72]	B/32	224	-	41.6	70.7	17.4	77.9	60.2	60.0	54.6
CE-CLIP [112]	B/32	224	-	40.7	76.0	34.8	86.0	55.7	57.0	58.4
NegCLIP [105]	B/32	224	-	42.4	73.7	30.5	83.6	65.0	62.5	59.6
READ-CLIP [46]	B/32	224	-	43.9	76.2	41.5	87.0	69.8	66.2	64.1
ReasonCLIP + READ [46]	B/32	224	-	45.2 ↑1.3	79.2 ↑3.0	44.9 ↑3.4	88.5 ↑1.5	69.2 ↓0.6	67.0 ↑0.8	65.7 ↑1.6
Eva-CLIP-02 [23]	L/14	224	2.0B	46.3	73.5	21.4	81.6	63.3	52.7	56.5
MetaCLIP [97]	L/14	224	2.5B	48.6	71.2	19.7	80.9	64.4	54.0	56.5
Long-CLIP [107]	L/14	224	0.4B	45.6	76.1	35.1	82.9	59.8	48.4	58.0
CLIP [69]	L/14	224	0.4B	40.7	68.8	20.5	73.4	60.6	44.3	51.4
ReasonCLIP	L/14	224	**+58M**	46.0 ↑5.3	75.9 ↑7.1	24.8 ↑4.3	78.2 ↑4.8	63.1 ↑2.5	59.5 ↑15.2	57.9 ↑6.5
SigLIP2 [86]	L/16	384	10.0B	45.5	70.6	15.9	81.3	65.7	51.7	55.1
CLIP [69]	L/14	336	0.4B	41.8	68.8	20.9	74.8	60.6	44.1	51.8
ReasonCLIP	L/14	336	**+58M**	46.3 ↑5.5	76.3 ↑7.5	24.8 ↑3.9	78.1 ↑3.3	63.8 ↑3.2	59.4 ↑15.3	58.1 ↑7.7
SigLIP [106]	So/14	384	10.0B	47.6	72.0	18.0	83.0	67.9	51.2	56.6
ReasonSigLIP	So/14	384	**+58M**	49.7 ↑2.1	76.3 ↑4.3	20.1 ↑2.1	84.1 ↑1.1	73.1 ↑5.2	72.9 ↑21.7	62.7 ↑6.1
SigLIP2 [86]	G/16	384	10.0B	46.9	71.4	16.5	83.0	67.4	48.8	55.7
ReasonSigLIP2	G/16	384	**+58M**	45.3 ↓1.6	77.9 ↑6.5	20.5 ↑4.0	76.8 ↓6.2	67.0 ↓0.4	66.9 ↑18.1	60.2 ↑4.5

Table 6: MLLM results on LLaVA-NeXT.

Model	AI2D [42]	ChartQA [62]	SciQA [73]	RealWordQA [93]	VisualLogic [98]	OKVQA [61]	GQA [33]	MME [25]	MMStar [12]	MMVP [85]
LLaVA-NeXT [57]	58.1	26.8	**73.1**	49.4	25.7	34.0	52.8	1296.1	39.7	58.7
w/ReasonCLIP	**58.6**	**28.0**	73.0	**49.5**	25.7	**35.8**	**53.4**	**1301.8**	**39.9**	**62.3**

tures and model scales, ReasonCLIP and ReasonSigLIP consistently outperform their base models on nearly all benchmarks, with average gains of approximately +5 to +7 points, demonstrating that large-scale reasoning supervision stably enhances structured compositional reasoning. Compared to data-centric approaches that rely on larger or higher-quality pretraining corpora, ReasonCLIP still achieves superior performance under the same backbone. Moreover, ReasonCLIP achieves competitive performance against specialized compositional methods without task-specific fine-tuning. Furthermore, applying READ [46] on top of ReasonCLIP-B/32 yields additional gains, validating its effectiveness as a stronger backbone for specialized compositional methods.

Observation 4. *ReasonCLIP as a visual tower unlocks stronger MLLM reasoning ability.* We replace CLIP with ReasonCLIP in LLaVA-NeXT [57] under an identical training setup and evaluate the model on commonsense-oriented MLLM benchmarks. Tab. 6 shows that, across multiple visual reasoning benchmarks, replacing CLIP with ReasonCLIP leads to consistent improvements, particularly on OKVQA, VisualLogic, GQA, and MMStar. This indicates that, within the same multimodal framework, introducing visual representations with stronger reasoning awareness can bring reproducible performance gains for downstream multimodal models. Meanwhile, we also conduct additional evaluations on sev-

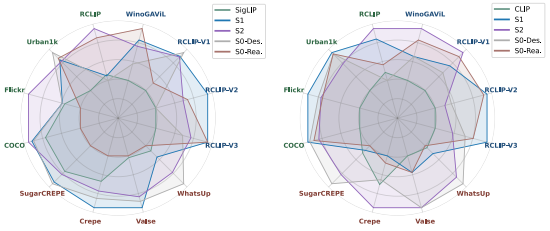

Fig. 4: Comparison of different training strategies.

Table 7: Ablation study.

Method	Retrieval		RCLIP-Bench		
	COCO	Flickr	V1	V2	V3
Stage 1	55.6	**50.5**	**26.6**	**19.2**	**24.3**
Des.-heavy	**55.8**	48.6	26.1	18.4	24.1
Rea.-heavy	52.5	46.0	23.5	15.3	19.0
Exp-Sched	53.2	47.7	25.8	17.6	23.3
Stage 2	53.2	**47.3**	**28.6**	**18.3**	**23.1**
w/o cls	**53.3**	47.3	**28.6**	18.0	22.6
w/ L2	53.1	47.1	28.4	17.3	23.0
Single-Label	52.8	46.5	28.1	17.5	22.0

eral non-reasoning benchmarks (e.g., AI2D and SciQA), and the results demonstrate that the model's performance on these non-reasoning tasks remains stable.

5.3 Ablation Study

Training Strategy. In Fig. 4, we present radar plots for ReasonCLIP-L14-336 and ReasonSigLIP-So/14-384 under different training strategies, comparing their performance across three evaluation dimensions.

For ReasonSigLIP, we observe that Stage 2 (S2) achieves the most balanced overall performance, particularly demonstrating stable advantages in retrieval tasks. However, on certain compositional benchmarks and RCLIP-Bench, Stage 1 (S1) performs better. This suggests that in scenarios requiring visually grounded reasoning, the progressive reasoning-aware alignment in S1 is more effective at preserving the synergy between visual semantic structure and reasoning signals, whereas explicit reasoning supervision in S2 may partially disrupt this balance.

For ReasonCLIP, the improvements from S1 are more pronounced, whereas those from S2 are relatively limited. This suggests that for CLIP backbones with limited capacity, reasoning enhancement must rely on strong visual grounding; excessive explicit supervision (e.g., S2) may weaken alignment. It also highlights the importance of large-scale pretraining and sufficient model capacity for effective reasoning improvement.

Training Configurations. Table 7 presents the ablation results under different training configurations, Des.-heavy / Rea.-heavy denote different branch weightings in Stage 1; Exp-Sched uses exponential scheduling; w/o cls removes category supervision; w/ L2 adds regularization; Single-Label replaces multi-label classification. The results show that our setting achieves the best overall performance, validating the effectiveness of our selected hyperparameters.

6 Limitation and Outlook

Limitations. Due to computational constraints, our 58M-training is smaller than industrial CLIP pretraining. Our contribution is to validate a scalable reasoning-oriented paradigm rather than compete with ultra-large-scale settings.

While evaluations are extensive, full coverage of CLIP applications is beyond scope; we release our models for broader community assessment.

Outlook. In this paper, we revisit CLIP-style encoders and investigate whether visually grounded reasoning can be incorporated through pretraining alone. Across architectures and scales, the proposed ReasonCLIP-58M framework enhances visually grounded reasoning and integrates seamlessly into MLLMs. These results demonstrate that structured reasoning supervision can systematically extend the representational capacity of CLIP-style encoders, providing a practical pathway toward reasoning-aware vision-language foundation models.

Acknowledgments and Disclosure of Funding

This research is funded by Khalifa University of Science and Technology through the Faculty Start-Ups under Project ID: KU-INT-FSU-2005-8474000775.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Alhamoud, K., Alshammari, S., Tian, Y., Li, G., Torr, P.H., Kim, Y., Ghassemi, M.: Vision-language models do not understand negation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29612–29622 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)
6. Basu, S., Hu, S.X., Sanjabi, M., Massiceti, D., Feizi, S.: Distilling knowledge from text-to-image generative models improves visio-linguistic reasoning in clip. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 6105–6113 (2024)
7. Bitton, Y., Bitton Guetta, N., Yosef, R., Elovici, Y., Bansal, M., Stanovsky, G., Schwartz, R.: Winogavil: Gamified association benchmark to challenge vision-and-language models. Advances in Neural Information Processing Systems **35**, 26549–26564 (2022)

8. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H.A., et al.: Perception encoder: The best visual embeddings are not at the output of the network. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
10. Chen, J., Yu, Q., Shen, X., Yuille, A., Chen, L.C.: Vitamin: Design scalable vision models in the vision-language era. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
11. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
12. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
13. Chen, Z., Liu, G., Zhang, B.W., Yang, Q., Wu, L.: Altclip: Altering the language encoder in clip for extended language capabilities. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8666–8682 (2023)
14. Cherti, M., Beaumont, R.: Clip benchmark (Nov 2022). `https://doi.org/10.5281/zenodo.15403103`, `https://doi.org/10.5281/zenodo.15403103`, accessed 27 June 2026
15. Chuang, Y.S., Li, Y., Wang, D., Yeh, C.F., Lyu, K., Raghavendra, R., Glass, J., Huang, L., Weston, J., Zettlemoyer, L., et al.: Meta clip 2: A worldwide scaling recipe. arXiv preprint arXiv:2507.22062 (2025)
16. Cui, W., Bi, K., Guo, J., Cheng, X.: More: Multi-modal retrieval augmented generative commonsense reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 1178–1192 (2024)
17. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (ICLR) (2024)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
19. Dumpala, S.H., Jaiswal, A., Shama Sastry, C., Milios, E., Oore, S., Sajjad, H.: Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. Advances in Neural Information Processing Systems **37**, 17972–18018 (2024)
20. El-Nouby, A., Klein, M., Zhai, S., Bautista, M.A., Toshev, A., Shankar, V., Susskind, J.M., Joulin, A.: Scalable pre-training of large autoregressive image models. arXiv preprint arXiv:2401.08541 (2024)
21. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
22. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
23. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing p. 105171 (2024)

24. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
25. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. Advances in Neural Information Processing Systems **38** (2026)
26. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)
27. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: Cyclip: Cyclic contrastive language-image pretraining. Advances in Neural Information Processing Systems **35**, 6704–6719 (2022)
28. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021)
29. Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., Globerson, A.: Incorporating structured representations into pretrained vision & language models using scene graphs. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 14077–14098 (2023)
30. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems **36**, 31096–31116 (2023)
31. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
32. Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., Dai, Q., Wang, C., Dai, X., Chen, D., et al.: Llm2clip: Powerful language model unlocks richer visual representation. arXiv preprint arXiv:2411.04997 (2024)
33. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
34. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, accessed 27 June 2026
35. Jampani, V., Maninis, K.K., Engelhardt, A., Truong, K., Karpur, A., Sargent, K., Popov, S., Araujo, A., Martin-Brualla, R., Patel, K., Vlasic, D., Ferrari, V., Makadia, A., Liu, C., Li, Y., Zhou, H.: NAVI: Category-agnostic image collections with high-quality 3d shape and pose annotations. In: NeurIPS (2023), `https://navidataset.github.io/`, accessed 27 June 2026
36. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
37. Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., Zhang, B., Fu, C., Gao, P., Li, H.: MME-CoT: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F.,

Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 27793–27830. PMLR (13–19 Jul 2025), https://proceedings.mlr.press/v267/jiang25n.html, accessed 27 June 2026
38. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
39. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
40. Kamath, A., Hessel, J., Chandu, K., Hwang, J.D., Chang, K.W., Krishna, R.: Scale can't overcome pragmatics: The impact of reporting bias on vision-language reasoning. arXiv preprint arXiv:2602.23351 (2026)
41. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? investigating their struggle with spatial reasoning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 9161–9175 (2023)
42. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
43. Koukounas, A., Mastrapas, G., Günther, M., Wang, B., Martens, S., Mohr, I., Sturua, S., Akram, M.K., Martínez, J.F., Ognawala, S., et al.: Jina clip: Your clip model is also your text retriever. arXiv preprint arXiv:2405.20204 (2024)
44. Koukounas, A., Mastrapas, G., Wang, B., Akram, M.K., Eslami, S., Günther, M., Mohr, I., Sturua, S., Martens, S., Wang, N., Xiao, H.: jina-clip-v2: Multilingual multimodal embeddings for text and images (2024), https://arxiv.org/abs/2412.08802, accessed 27 June 2026
45. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
46. Kwon, J., Min, K., Sohn, J.y.: Enhancing compositional reasoning in clip via reconstruction and alignment of text descriptions. arXiv preprint arXiv:2510.16540 (2025)
47. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
48. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023)
49. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
50. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
51. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)

52. Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., Chang, S.F.: Clip-event: Connecting text and images with event structures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16420–16429 (2022)
53. Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al.: What if we recaption billions of web images with llama-3? arXiv preprint arXiv:2406.08478 (2024)
54. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021)
55. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23390–23400 (2023)
56. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
57. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, accessed 27 June 2026
58. Liu, Y., Li, X., Wang, Z., Zhao, B., Xie, C.: Clips: An enhanced clip framework for learning with synthetic captions. arXiv preprint arXiv:2411.16828 (2024)
59. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
60. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
61. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
62. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
63. McKinzie, B., Gan, Z., Fauconnier, J.P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Belyi, A., et al.: Mm1: methods, analysis and insights from multimodal llm pre-training. In: European Conference on Computer Vision. pp. 304–323. Springer (2024)
64. Oh, Y., Cho, J.W., Kim, D.J., Kweon, I.S., Kim, J.: Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality. arXiv preprint arXiv:2410.05210 (2024)
65. Onoe, Y., Rane, S., Berger, Z., Bitton, Y., Cho, J., Garg, R., Ku, A., Parekh, Z., Pont-Tuset, J., Tanzer, G., et al.: Docci: Descriptions of connected and contrasting images. In: European Conference on Computer Vision. pp. 291–309. Springer (2024)
66. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

67. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., Gatt, A.: Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8253–8280 (2022)
68. Patel, M., Kusumba, N.S.A., Cheng, S., Kim, C., Gokhale, T., Baral, C., et al.: Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. Advances in neural information processing systems **37**, 32731–32760 (2024)
69. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
70. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18082–18091 (2022)
71. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
72. Sahin, U., Li, H., Khan, Q., Cremers, D., Tresp, V.: Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5563–5573 (2024)
73. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022)
74. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
75. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
76. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
77. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. arXiv preprint arXiv:2103.01913 (2021)
78. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5198–5215 (2022)
79. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
80. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024)
81. Team, Q.: Qwen3 technical report (2025), `https://arxiv.org/abs/2505.09388`, accessed 27 June 2026

82. Thapliyal, A.V., Tuset, J.P., Chen, X., Soricut, R.: Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 715–729 (2022)
83. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
84. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
85. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
86. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
87. Wan, Y., Ma, Y., You, H., Wang, Z., Chang, S.F.: Bridging the gap between recognition-level pre-training and commonsensical vision-language tasks. In: Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022). pp. 23–35 (2022)
88. Wang, B., Ning, Z., Ding, J., Gao, X., Li, Y., Jiang, D., Yang, J., Liu, W.: Fixclip: Dual-branch hierarchical contrastive learning via synthetic captions for better understanding of long text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20694–20704 (2025)
89. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. In: European conference on computer vision. pp. 315–332. Springer (2024)
90. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
91. Wang, Z., Yang, S., Qiao, L., Ma, L.: Vitrix-clipin: Enhancing fine-grained visual understanding in clip via instruction editing data and long captions. arXiv preprint arXiv:2508.02329 (2025)
92. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
93. xAI: Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v` (Apr 2024), `https://x.ai/news/grok-1.5v`, accessed 27 June 2026
94. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010)
95. Xie, C., Wang, B., Kong, F., Li, J., Liang, D., Zhang, G., Leng, D., Yin, Y.: Fg-clip: Fine-grained visual and textual alignment. arXiv preprint arXiv:2505.05071 (2025)
96. Xu, G., Jin, P., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440 (2024)
97. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)

98. Xu, W., Wang, J., Wang, W., Chen, Z., Zhou, W., Yang, A., Lu, L., Li, H., Wang, X., Zhu, X., et al.: Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. arXiv preprint arXiv:2504.15279 (2025)
99. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29968–29978 (2025)
100. Ye, S., Xie, Y., Chen, D., Xu, Y., Yuan, L., Zhu, C., Liao, J.: Improving commonsense in vision-language models via knowledge graph riddles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2634–2645 (2023)
101. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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 (2014). https://doi.org/10.1162/tacl_a_00166, https://aclanthology.org/Q14-1006/, accessed 27 June 2026
102. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
103. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
104. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)
105. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=KRLUvxh8uaX, accessed 27 June 2026
106. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
107. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: European conference on computer vision. pp. 310–325. Springer (2024)
108. Zhang, H., Ee, Y.K., Fernando, B.: Rca: Region conditioned adaptation for visual abductive reasoning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9455–9464 (2024)
109. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
110. Zhang, H., Ren, S., Yuan, H., Zhao, J., Li, F., Sun, S., Liang, Z., Yu, T., Shen, Q., Cao, X.: Mmvp: A multimodal mocap dataset with vision and pressure sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21842–21852 (2024)
111. Zhang, J., Qu, X., Zhu, T., Cheng, Y.: Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5406–5419 (2025)

112. Zhang, L., Awal, R., Agrawal, A.: Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13774–13784 (2024)
113. Zhao, W., Huang, Z., Feng, J., Wang, X.: Superclip: Clip with simple classification supervision. arXiv preprint arXiv:2512.14480 (2025)
114. Zhao, W., Huang, Z., Feng, J., Wang, X.: Superclip: clip with simple classification supervision. Advances in Neural Information Processing Systems **38**, 117828–117850 (2026)
115. Zheng, C., Zhang, J., Kembhavi, A., Krishna, R.: Iterated learning improves compositionality in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13785–13795 (2024)
116. Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., Shen, Y.: Dreamlip: Language-image pre-training with long captions. In: European Conference on Computer Vision. pp. 73–90. Springer (2024)
117. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
118. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)