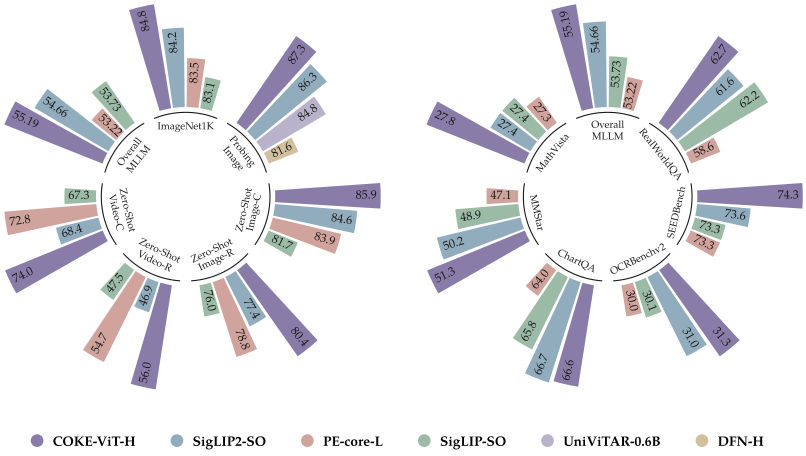
COKE: A UNIFIED IMAGE-VIDEO ENCODER TRAINED ON PUBLIC CORPUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Leading vision encoders often rely on training corpora that are not publicly released, limiting access to comparable training resources. We present **COKE**, a family of unified image-video encoders and a training recipe using public data and priors from released models. It includes COKE-ViT-H (0.6B) and COKE-ViT-G (1.9B), with shared image-video weights at each resolution. ***Visual Reference Pretrain*** applies auxiliary visual supervision through dedicated reference-token outputs and improves early sample efficiency. At a fixed distillation weight, this pathway achieves higher classification and retrieval scores than pooled-patch distillation. ***Snapshot-Replay*** periodically interpolates video-trained weights with a fixed image-pretrained checkpoint and resumes optimization from the fused weights. In controlled COKE-ViT-H-256 experiments, it improves image and video averages over continuous video-only fine-tuning by 1.81 and 1.03 points, respectively, without replaying image-text data. A resolution curriculum supports fixed and native inputs, while COKE-K and COKE-V guide concept balancing alongside quality filtering and deduplication to construct the 1.68B-pair **COKE-1.7B** corpus. Fixed-budget ablations support the effectiveness of COKE-K and COKE-V for vocabulary-based image-text data selection. At 512 resolution, COKE-CLIP-G achieves 86.8 average zero-shot accuracy across six image classification benchmarks and 79.9 average retrieval Recall@1, compared with 86.6 and 78.9 for PE_{core}-G under our evaluation protocol. Frozen-encoder probing and multimodal evaluations demonstrate the utility of the learned representations beyond zero-shot image-text tasks. We will release models, vocabularies, and evaluation code.



1 INTRODUCTION

Vision encoders support recognition, retrieval, and multimodal understanding, requiring semantic discrimination, language alignment, spatial detail, and temporal aggregation. Language alignment enables open-vocabulary prediction and cross-modal retrieval and provides visual inputs to multimodal large language models (Liu et al., 2024a; Chen et al., 2024c; Wang et al., 2024a). Contrastive image-text pretraining (Radford et al., 2021; Zhai et al., 2023) is widely adopted, but leading recipes demand substantial resources. SigLIP 2 (Tschannen et al., 2025) processes 40B samples from WebLI-10B,

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.

and Perception Encoder (Bolya et al., 2025) processes 86B from CC-5.4B, requiring curated data and substantial computation. Unifying these capabilities with accessible resources calls for effective curation and training. Native-resolution inputs (Dehghani et al., 2023b; Wang et al., 2024a; Qiao et al., 2025) and video understanding (Bolya et al., 2025) additionally require schedules that accommodate varying token counts across image sizes and frame counts while controlling computation.

We study competitive unified encoders using public training data and released model priors, motivated by three considerations. *(1) Sample-efficient contrastive learning.* Limited budgets motivate sample-efficient image-text contrastive learning. Released self-supervised visual encoders provide early learning priors (Oquab et al., 2023; Siméoni et al., 2025). We study auxiliary visual supervision for early sample efficiency while supporting classification and cross-modal retrieval. *(2) Unified image-video representation.* Unified encoders share representations across image and video understanding. Video adaptation can degrade pretrained image performance, motivating training that develops video capability while preserving image-task performance. *(3) Public-data curation.* Public web datasets contain uneven concept frequencies and variable image-text alignment. Model-based filtering (Fang et al., 2023) and metadata-based balancing (Xu et al., 2023) play complementary roles. MetaCLIP uses WordNet and Wikipedia metadata. Since vocabulary guides selection and balancing, we study how vocabulary choice and resulting data affect performance under a fixed budget.

We present **COKE**, named after the ***CO**mmon-**K**nowledg**E*** vocabulary used in its data pipeline, at two scales, COKE-ViT-H (0.6B) and COKE-ViT-G (1.9B), using public corpora and released model priors. Specifically, DINOv3 provides visual supervision, while SigLIP 2 supports text initialization and data filtering. The 1.68B-pair corpus excludes auxiliary model pretraining data. Figure 5 in the appendix compares zero-shot image classification against direct training-corpus size. At 512 resolution, COKE-CLIP-G achieves an average accuracy of 86.8 across six classification benchmarks, compared with 86.6 for PE_{core}-G. The recipe combines a unified encoder, two training strategies, a resolution schedule, and a data engine. ***(1) Visual Reference Pretrain.*** We introduce reference tokens *[REF]* whose outputs are mapped through a trainable student projector and supervised by features from a frozen visual teacher. Distillation uses the reference token output, while image-text contrastive learning uses pooled patch features. The two objectives update a shared student backbone. Under the same distillation weight and training settings, the tested reference-token design yields better classification and retrieval scores than pooled-patch distillation. The teacher branch and distillation loss are removed after this stage, while the reference tokens remain in the student. ***(2) Snapshot-Replay.*** Both modalities share a spatiotemporal patchifier and encoder. Replacing UniViTAR's alternating image-video updates, each resolution uses a video-only branch initialized from its image checkpoint. After each short burst, weights are interpolated with that fixed checkpoint, and training resumes from the fused weights. With matched initialization and video-text data, Snapshot-Replay improves image and video performance over continuous video-only fine-tuning. ***(3) Resolution curriculum.*** Training proceeds through 256, 384, 512, and native resolution, concentrating image-text samples at 256 to control computation and ultimately supporting variable image sizes and aspect ratios. ***(4) Common-knowledge data engine.*** **COKE-K** contains approximately 1.1M entries derived from a consolidated commonsense knowledge graph (Ilievski et al., 2021; Speer et al., 2017). **COKE-V** contains approximately 21K visual category entries from ImageNet-21K (Deng et al., 2009). Quality filtering, vocabulary-based balancing, and deduplication produce the **COKE-1.7B** corpus of 1.68B image-text pairs. In COKE-ViT-H-256 ablations without VRP, replacing the MetaCLIP vocabulary with COKE-K on the same candidate pool improves the combined classification and retrieval average by 0.80 percentage points. COKE-V selection from the public pool improves the same average by 2.48 points over MetaCLIP selection under a fixed cumulative 1B-sample budget (§3.3.3).

Our contributions are summarized as follows.

- (**1**) We develop COKE, a two-scale family of unified image-video encoders using public corpora and released model priors. Its recipe supports fixed and native resolutions and is competitive in zero-shot classification, retrieval, frozen probing, and multimodal understanding.
- (**2**) We present Visual Reference Pretrain for teacher supervision of reference-token outputs and pooled-patch text alignment, and Snapshot-Replay for resuming video fine-tuning after each interpolation toward a fixed image-pretrained checkpoint.
- (**3**) We build COKE-K and COKE-V from public knowledge and COKE-1.7B via vocabulary-based balancing, quality filtering, and deduplication. In fixed-budget experiments, COKE-K has the highest retrieval average, while COKE-V leads in image classification.

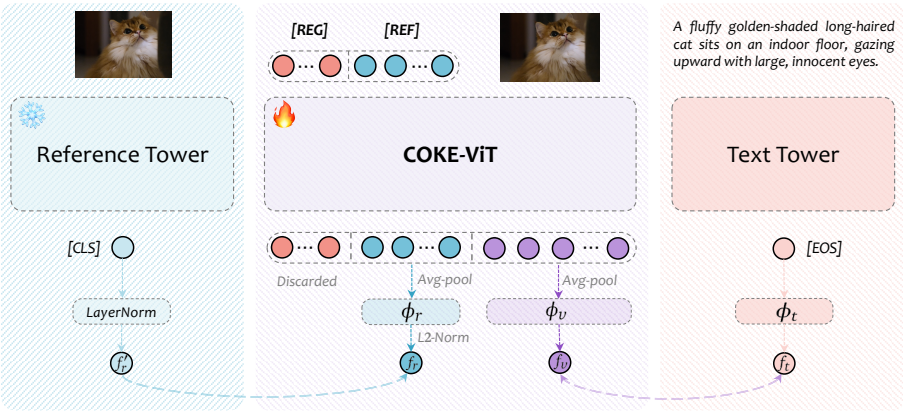


Figure 1: **Overview of COKE-CLIP.** The student processes patch, register *[REG]*, and reference *[REF]* tokens. Patch features are pooled and projected by ϕ_v to form f_v, which is aligned with text embedding f_t. Reference-token outputs are averaged, projected by ϕ_r, and L2-normalized to form f_r. After VRP, the teacher branch is removed and the student's *[REF]* tokens remain.

2 COKE

2.1 Unified Image-Video Architecture

COKE-CLIP combines a randomly initialized COKE-ViT with a text tower initialized from a released SigLIP 2 checkpoint (Tschannen et al., 2025), as shown in Fig. 1. The visual encoder supports fixed-resolution inputs and variable-length native-resolution sequences, following prior native-resolution encoders (Dehghani et al., 2023b; Wang et al., 2024a; Qiao et al., 2025). We instantiate COKE-ViT-H and COKE-ViT-G with approximately 0.6B and 1.9B parameters. Their configurations are provided in Appendix B.1.

Shared visual representation. Given T sampled video frames of size $H \times W$, consecutive frames are grouped into $T' = T/p_t$ temporal cubes with $p_t = 2$ frames each. Each cube is partitioned into non-overlapping spatiotemporal patches of size (p_t, p_s, p_s), where p_s is the spatial patch size. A shared linear embedding maps each patch to a d-dimensional token, yielding $H'W'$ tokens per cube, with $H' = H/p_s$ and $W' = W/p_s$. Duplicating an image into a two-frame cube allows both modalities to use the same patchifier and backbone. Cubes with different spatial token grids are packed without padding, and a block-diagonal mask restricts attention to each cube, including cubes from the same video. Patch features are averaged within each cube and then across cubes to form a video descriptor. This descriptor is invariant to cube permutations, although each two-frame cube can encode temporal information. Native-resolution inputs approximately preserve aspect ratios, with spatial dimensions adjusted to the token budget. RoPE-2D uses spatial coordinates rather than a learned positional table and applies only to patch tokens. Register and reference tokens interact with patch tokens without rotary positional encoding. Following register tokens (Darcet et al., 2024) and dedicated distillation tokens (Touvron et al., 2021), both scales append four register tokens and eight reference tokens, denoted *[REG]* and *[REF]*, without a *[CLS]* token. These non-patch tokens participate in every self-attention layer. Reference-token outputs provide the visual supervision interface described in §2.2, while pooled patch features support image-text alignment. Register and reference tokens are excluded from the patch-feature pool.

Training schedule. Training processes 28.8B image-text and 2.56B video-text samples across 256, 384, 512, and native resolution. At R=256, VRP is followed by image-text contrastive training without the teacher. Each image-text checkpoint initializes the next resolution level and an independent video-text branch trained with Snapshot-Replay. Video branches produce shared image-video weights at their respective resolutions and do not initialize subsequent image-training levels. Most image-text samples are allocated to the lowest resolution to control computation before adaptation to higher spatial detail and variable aspect ratios. The native stage limits each sample to 16384 patch tokens before patch dropout, resizing inputs that exceed this budget. The curriculum is part of the complete recipe rather than an isolated schedule ablation. Appendix B.2 provides the training diagram and resolution details, and Table 8 reports the sample allocation.

2.2 VISUAL REFERENCE PRETRAIN

Visual Reference Pretrain applies auxiliary supervision from a frozen DINOv3-7B teacher (Siméoni et al., 2025) during the first 8B image-text samples at R=256. Teacher matching uses reference-token outputs, while language alignment uses pooled patch features. This assigns the objectives to distinct outputs of a shared backbone without directly imposing distillation on the pooled patch representation. Both losses update the shared backbone, so teacher supervision can influence the contrastive representation through shared parameters and token interactions. The design does not imply independent optimization of the two objectives.

Feature definitions. The pooled patch descriptor is mapped through ϕ_v, a LayerNorm followed by a linear projection, to obtain the visual embedding f_v. The last-layer *EOS* state of the text tower is projected by ϕ_t to obtain f_t. Both embeddings are L2-normalized for contrastive learning. The pretrained text tower is frozen during VRP and updated in later image-text and video-text training. Reference-token outputs are averaged and passed through a separate, newly initialized student projector ϕ_r, also consisting of a LayerNorm and a linear projection, followed by L2 normalization to obtain f_r. The teacher target f'_r is the DINOv3 *CLS* feature after its output LayerNorm, without an additional projection head or L2 normalization. Teacher features are computed without gradient tracking. The reference embedding matches the teacher feature dimension, while the visual embedding matches the text embedding dimension. Register tokens are excluded from both student output pools.

Joint objective. Image-text alignment uses the sigmoid contrastive loss (Tschannen et al., 2025), and visual distillation minimizes the Kullback-Leibler divergence between teacher and student feature distributions. The combined objective is

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\sigma\big(y_{ij}(\tau s_{ij}+b)\big) + \lambda\,\mathcal{D}_{\mathrm{KL}}(q_t \,\|\, q_s), \tag{1}$$

where N counts image-text pairs across all devices and $s_{ij} = \langle f_v^{(i)}, f_t^{(j)}\rangle$ is their cosine similarity. The label y_{ij} is $+1$ for matched pairs and -1 otherwise. The logit scale τ and bias b are learnable, and λ weights distillation. The sigmoid term sums over all pairs and is normalized by the image count. The feature distributions are

$$q_t = \mathrm{softmax}(f'_r/T_t), \qquad q_s = \mathrm{softmax}(f_r/T_s), \tag{2}$$

with softmax applied along the feature dimension and fixed temperatures $T_t = 1$ and $T_s = 0.1$. KL sums over feature dimensions and averages samples per device, with no extra temperature-squared factor. It updates ϕ_r and the shared student backbone without directly distilling pooled patch outputs.

Training stages and empirical support. The auxiliary objective is limited to early training to support initial representation learning before continued optimization for image-text alignment. After VRP, the teacher and distillation loss are removed, while all eight reference tokens remain active during subsequent training and inference. The student retains the same token configuration and continues image-text alignment without teacher computation. Under a fixed distillation weight and training configuration, reference-token supervision achieves higher classification and retrieval scores than pooled patch distillation in Table 4(b). A separate controlled COKE-ViT-SO400M-384 experiment compares VRP with a baseline using the same token configuration and optimization settings. Both runs use randomly initialized visual backbones and frozen SigLIP 2 text towers. The VRP run reaches 76.4 mean zero-shot classification accuracy over six benchmarks after 0.6B samples, exceeding the baseline's 75.5 after 1.6B. At 1.6B samples, VRP obtains 82.6. The approximately 2.7× reduction concerns sample count and excludes teacher computation, without implying a wall-clock or FLOP speedup. This configuration is specific to the convergence experiment, with teacher supervision throughout the VRP run. Appendix D.2 provides the curves, complete setup, and external reference.

2.3 SNAPSHOT-REPLAY FINETUNE

The shared architecture allows video adaptation without an extra temporal module. At each resolution, a video branch starts from its image-text checkpoint and uses only video-text pairs. The video budget is split equally among four levels, at 640M samples each. Continuous video fine-tuning may reduce image performance under this distribution shift. Snapshot-Replay periodically interpolates video-trained parameters with the fixed image checkpoint and resumes training from the blend. Its effect on image retention and video adaptation is tested with matched initialization and video data in §3.3.2.

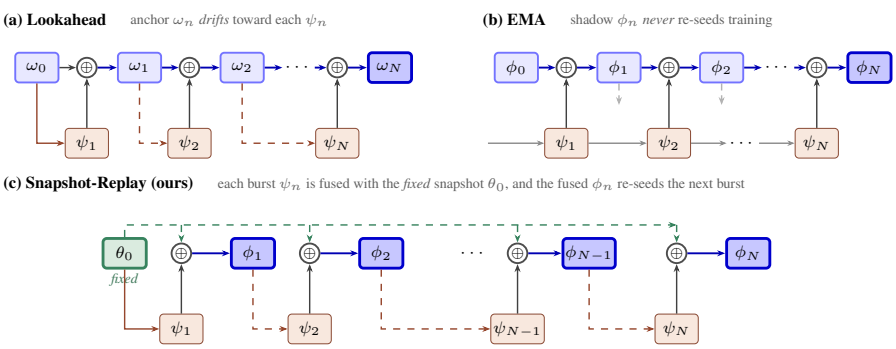


Figure 2: **Snapshot-Replay versus Lookahead and EMA.** Red nodes mark training endpoints ψ_n, blue nodes slow or fused weights, and $\oplus$ interpolation. (a) Lookahead resumes training from weights blended with a moving target. (b) Evaluation-only EMA never feeds averaged parameters back into training. (c) Snapshot-Replay uses fixed θ_0, resumes from each fused ϕ_n, and deploys ϕ_N.

Periodic interpolation and resumed optimization. Let θ_0 be the image-text checkpoint at the current resolution, held fixed throughout its video branch. Each round performs K optimization steps from the preceding fused parameters and interpolates the resulting endpoint with θ_0

$$\psi_n = \mathrm{Train}_K(\phi_{n-1}), \qquad \phi_n = \alpha\theta_0 + (1-\alpha)\psi_n, \qquad \phi_0 = \theta_0, \tag{3}$$

where $\alpha \in [0, 1]$ controls the contribution of the image-pretrained checkpoint. Optimization resumes from ϕ_n after every interpolation, and the final fused checkpoint ϕ_N is used after N rounds. Setting $\alpha = 0$ yields continuous video-text fine-tuning, while $\alpha = 1$ recovers the initial checkpoint. The interpolation admits an equivalent regularized minimization, and a bounded-update assumption yields a bound on parameter displacement from θ_0, as detailed in Appendix D.3. This bound does not guarantee image-performance retention, which is assessed empirically.

Relation to weight averaging. Fig. 2 compares the interpolation targets and use of interpolated parameters in subsequent optimization. Lookahead (Zhang et al., 2019) updates its slow parameters using $\omega_n = (1-\gamma)\omega_{n-1} + \gamma\psi_n$, with interpolation coefficient γ, and resumes training from them, so its target evolves. Snapshot-Replay instead uses the fixed θ_0. Evaluation-only EMA maintains averaged parameters without feeding them back into training. WiSE-FT (Wortsman et al., 2022b) interpolates pretrained and fine-tuned parameters after fine-tuning. In Snapshot-Replay, each periodic interpolation influences subsequent optimization.

2.4 Common-Knowledge based Data Curation

We build **COKE-1.7B** from public datasets via quality filtering, concept-based balancing, and deduplication. Vocabulary guides data selection, complementing image-text similarity filtering (Xu et al., 2023). We use released auxiliary models, with the resource scope given in Appendix D.6.

Knowledge vocabularies. COKE-K contains 1,099,017 uni-gram and bi-gram entries from CSKG (Ilievski et al., 2021), which consolidates resources including ConceptNet (Speer et al., 2017), Wiktionary (Meyer & Gurevych, 2012), and WordNet (Miller, 1995). COKE-V provides approximately 21K visual category names from ImageNet-21K (Deng et al., 2009). Both vocabularies supply textual entries for caption matching and subsequent sampling. Vocabulary size counts available entries, while actual concept coverage depends on caption occurrences and retention during curation.

Concept-based selection. We aggregate DataComp-XL (Gadre et al., 2023), LAION-5B (Schuhmann et al., 2022), and COYO-700M (Byeon et al., 2022). Captions from multilingual LAION subsets are translated into English to match the same concept vocabularies. Image-validity screening, SigLIP 2 similarity filtering, and deduplication produce a 4.5B-pair candidate pool. Text preprocessing and vocabulary matching follow MetaCLIP. For each concept, all matching pairs are retained up to a specified cap. Above the cap, pairs are sampled uniformly without replacement. Selections are merged and deduplicated per pass. As captions may match multiple entries, the cap constrains each concept-specific selection before merging rather than its frequency in the merged dataset.

Selection paths and final corpus. Two independent passes apply COKE-K and COKE-V to the 4.5B pool, each capped at 60K pairs per concept, yielding 1.42B and 0.40B pairs, respectively. A

third applies COKE-K directly to the original, already score-filtered DataComp-1B, bypassing score filtering and yielding 0.39B pairs at a 40K cap. Each pass limits frequent concepts' contributions and retains its vocabulary-based selection separately until merging. Despite deduplication within each pass, selections overlap. Global deduplication using URLs and image-content hashes reduces their combined 2.21B pairs by approximately 24% to 1.68B, forming COKE-1.7B. This covers only the training corpus, without additional benchmark-specific decontamination. Each training pass draws 1.6B pairs, so 28.8B image-text samples correspond to 18 passes. Appendix B.3 details translation, score-filtering thresholds, intermediate pool sizes, and the full pipeline. Section 3.3.3 compares COKE-K and COKE-V selection with MetaCLIP on the same public pool at a fixed sample budget.

3 Experiments

3.1 Experimental Setup

The main models process 28.8B image-text samples from COKE-1.7B and 2.56B video-text samples. The video corpus contains approximately 64M valid pairs from Panda-70M (Chen et al., 2024b), WebVid-10M (Bain et al., 2021), and InternVid-10M-FLT (Wang et al., 2023), retaining released captions. Image epochs draw 1.6B pairs, and each resolution uses 640M video-text samples over 10 passes. Each clip's 16 sampled frames form eight two-frame cubes. Training uses AdamW with cosine decay, default VRP weight $\lambda = 4000$, and Snapshot-Replay ($K = 10$, $\alpha = 0.75$). Appendices C.1.1 and C.1.2 detail data processing, sample allocation, and the full training configuration. Zero-shot evaluation reports top-1 classification accuracy and bidirectional retrieval Recall@1. For models evaluated here, we report the higher center crop or squash resize score per dataset and apply PE retrieval reweighting. Classification averages task-specific prompt embeddings, whereas retrieval uses unprompted captions. Tables distinguish re-evaluated baselines from published results. Frozen-encoder probing fits attentive heads on ImageNet-1K, while multimodal instruction tuning updates the vision tower. Appendix C provides complete protocols and per-benchmark results.

3.2 Evaluation of the Complete COKE Models

3.2.1 Zero-shot Image Classification and Retrieval

Table 1 reports means over six classification benchmarks and four retrieval directions. At 384, COKE-H obtains 85.9 and 80.4, compared with 84.6 and 77.4 for SigLIP 2-SO. H's largest classification gain over SigLIP 2-SO occurs on ImageNet-A (92.1 versus 87.0), with improvements across all four retrieval directions. At 512, COKE-G obtains 86.8 and 79.9, compared with 86.6 and 78.9 for PE-G. Native inputs yield the highest COKE retrieval averages, while classification is 0.7 points below the corresponding 512 checkpoints. Across resolutions, G improves classification over H but scores lower on retrieval. These progressively trained checkpoints also differ in training duration. Corpus-size comparisons concern direct training data and exclude auxiliary-model pretraining, so the results assess the complete recipe rather than isolate corpus size.

3.2.2 Zero-shot Video Classification and Retrieval

Table 2 averages three classification benchmarks and six retrieval directions. At $384/16$, COKE-H scores 74.0 and 56.0, exceeding PE-L at $336/14$ by 1.2 and 1.3 points. H improves Kinetics-400 and UCF-101 and bidirectional retrieval on ActivityNet and MSVD, but trails on HMDB-51 and MSR-VTT. Both encode 576 spatial patch tokens per unit, but COKE uses eight two-frame cubes versus PE's eight frames. Matched patch counts imply neither equal frame counts nor computation. At $512/16$, COKE-G scores 75.3 and 55.7, below PE-G's 75.9 and 58.5. Higher resolution favors COKE-G classification and lower resolution retrieval. Native H and G outperform their UniViTAR counterparts by 4.9 and 5.7 classification and 2.0 and 2.8 retrieval points.

3.2.3 Frozen-Encoder Probing

With frozen visual backbones and attentive heads, COKE-H and COKE-G obtain averages of 87.3 and 87.9 at 576 patch tokens, compared with 86.3 and 87.6 for SigLIP 2-SO and SigLIP 2-g-1B. Against their SigLIP 2 counterparts, H and G improve most on ImageNet-A (5.3 and 2.9 points) but trail on ImageNet-Sketch (1.6 and 2.1). At 1024 patch tokens, the COKE averages reach 87.5 and 88.0. Table 12 in the appendix lists all probing results, with linear-head baselines for reference.

Table 1: **Zero-shot image classification and retrieval.** 👁/🙈 mark public/nonpublic data sources, and † our re-evaluation. **Overall** averages six classification benchmarks or four retrieval directions equally. Blocks group fixed/native inputs and $\leq$1B/$>$1B models. Blockwise column best/second-best are **bold**/underlined, with COKE-CLIP shaded. Full results and sources appear in Table 10.

Method	Data Source	Res./Patch	Overall	Classification						Overall	Flickr		COCO	
				IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net		T→I	I→T	T→I	I→T
SigLIP-SO (Zhai et al., 2023)	🙈 WebLI-10B	384/16	81.7	83.1	82.5	95.8	77.2	74.5	77.0	76.0	83.0	94.3	54.2	72.4
†SigLIP2-SO (Tschannen et al., 2025)	🙈 WebLI-10B	384/16	84.6	84.2	87.0	96.1	78.5	**76.2**	85.4	77.4	84.1	94.9	56.9	73.7
†PE_{core}-L (Bolya et al., 2025)	🙈 CC-5.4B	336/14	83.9	83.5	89.1	95.2	77.9	73.4	84.5	78.8	85.5	96.5	57.1	75.9
COKE-CLIP-H	👁 COKE-1.7B	256/16	85.0	84.4	89.6	96.5	79.0	75.7	84.7	80.1	86.4	96.6	59.4	77.9
COKE-CLIP-H	👁 COKE-1.7B	384/16	85.9	84.8	92.1	**96.7**	**79.8**	76.0	86.0	**80.4**	**87.1**	96.6	59.6	**78.1**
COKE-CLIP-H	👁 COKE-1.7B	512/16	**86.1**	**84.9**	**92.7**	**96.7**	79.6	76.1	**86.4**	**80.4**	86.8	**96.7**	**60.1**	78.0
UniViTAR-H (Qiao et al., 2025)	👁 Merged-1B	Native/14	82.1	82.3	86.8	94.9	76.1	71.6	81.1	76.6	84.1	95.5	55.4	71.7
Qwen3-ViT-SO (Bai et al., 2025)	🙈 In-House	Native/16	83.6	84.6	87.1	95.7	78.8	74.5	81.0	-	-	-	-	-
COKE-CLIP-H	👁 COKE-1.7B	Native/16	**85.4**	**84.8**	**91.6**	**96.7**	**79.5**	**75.9**	**83.8**	**80.8**	**87.4**	**96.9**	**60.8**	**78.0**
CLIPAv2-G (Li et al., 2023d)	👁 DC-1B	336/14	82.7	83.1	86.0	95.4	77.3	74.5	79.7	72.2	78.3	92.2	50.4	67.8
EVA-8B (Sun et al., 2023a)	👁 Merged-2B	224/14	82.9	83.5	85.2	95.3	77.7	74.3	81.2	74.9	80.8	95.6	53.0	70.3
†SigLIP2-g-1B (Tschannen et al., 2025)	🙈 WebLI-10B	384/16	86.2	85.0	90.5	96.6	80.0	**77.4**	87.5	78.1	84.2	95.8	57.4	75.2
† PE_{core}-G (Bolya et al., 2025)	🙈 CC-5.4B	448/14	86.6	**85.4**	92.6	96.5	80.2	76.5	**88.2**	78.9	85.7	96.2	58.2	75.5
COKE-CLIP-G	👁 COKE-1.7B	256/16	85.8	84.9	91.3	96.5	79.5	76.3	86.1	79.6	**86.5**	96.3	58.6	76.9
COKE-CLIP-G	👁 COKE-1.7B	384/16	86.6	85.3	93.3	96.9	80.2	77.1	86.7	**79.9**	**86.5**	96.6	59.3	**77.1**
COKE-CLIP-G	👁 COKE-1.7B	512/16	**86.8**	85.3	**93.5**	**97.0**	**80.5**	77.2	87.1	**79.9**	86.4	**97.1**	**59.6**	76.5
UniViTAR-1B (Qiao et al., 2025)	👁 Merged-1B	Native/14	83.5	82.9	89.1	95.7	77.3	73.4	82.8	76.3	83.5	95.1	55.3	71.3
COKE-CLIP-G	👁 COKE-1.7B	Native/16	**86.1**	**85.1**	**93.1**	**96.9**	**80.0**	**77.1**	**84.3**	**80.5**	**87.2**	**96.8**	**60.0**	**77.8**

Table 2: **Zero-shot video classification and retrieval.** Baselines are quoted, except our † re-evaluations. **Overall** gives unweighted means over three classification (blue) or six retrieval (green) metrics. Grouping, boldface, and underlining follow Table 1. Gray rows denote video-specific models. **Frames** 16 (8) denotes eight cubes pairing 16 uniformly sampled frames consecutively.

Method	Res./Patch	Frames	Overall	Classification			Overall	ANet		MSR-VTT		MSVD	
				K400	UCF	HMDB		V→T	T→V	V→T	T→V	V→T	T→V
SigLIP-SO (Zhai et al., 2023)	384/16	16	67.3	66.8	83.0	52.1	47.5	36.6	39.3	37.5	41.1	75.5	54.7
†SigLIP2-SO (Tschannen et al., 2025)	384/16	8	68.4	67.0	87.4	50.9	46.9	34.9	35.8	36.0	42.5	76.9	55.3
† PE_{core}-L (Bolya et al., 2025)	336/14	8	72.8	72.9	87.1	**58.5**	54.7	42.8	45.4	**50.1**	**50.3**	82.5	57.1
COKE-CLIP-H	256/16	16 (8)	73.6	74.1	89.4	57.3	**56.6**	**47.4**	**49.1**	49.1	50.2	**84.2**	59.4
COKE-CLIP-H	384/16	16 (8)	**74.0**	**74.7**	89.8	57.6	56.0	47.1	47.8	48.4	49.8	83.4	**59.7**
COKE-CLIP-H	512/16	16 (8)	**74.0**	74.5	**89.9**	57.7	55.5	45.9	48.2	46.6	49.6	83.4	59.4
UniViTAR-0.6B (Qiao et al., 2025)	Native/14	2∼32	68.6	67.6	82.9	55.2	54.9	**48.7**	**51.5**	48.6	**50.2**	75.8	54.3
COKE-CLIP-H	Native/16	16 (8)	**73.5**	**74.0**	**88.9**	**57.5**	**56.9**	48.1	49.8	**50.1**	49.9	**84.0**	**59.6**
EVA01-g (Sun et al., 2023a)	224/14	16	62.8	63.4	72.1	52.9	45.5	37.0	40.1	37.2	40.1	67.6	50.8
InternViT-6B (Chen et al., 2024c)	224/14	8	-	69.1	-	-	-	-	-	42.4	46.3	-	-
† PE_{core}-G (Bolya et al., 2025)	448/14	8	**75.9**	**76.4**	90.2	**61.0**	**58.5**	**51.6**	**53.3**	**50.2**	**51.2**	**85.2**	59.7
COKE-CLIP-G	256/16	16 (8)	74.5	75.2	91.6	56.6	57.1	47.7	49.7	49.9	51.0	84.5	59.8
COKE-CLIP-G	384/16	16 (8)	75.0	75.9	91.9	57.3	56.7	47.9	49.9	48.3	50.1	83.4	**60.3**
COKE-CLIP-G	512/16	16 (8)	75.3	76.0	**92.4**	57.6	55.7	46.7	48.4	47.7	48.5	82.8	59.8
UniViTAR-1B (Qiao et al., 2025)	Native/14	2∼32	69.0	68.6	81.0	57.3	54.0	47.8	**49.6**	48.3	47.6	75.5	55.2
COKE-CLIP-G	Native/16	16 (8)	**74.7**	**75.2**	**91.1**	**57.8**	**56.8**	**48.1**	48.7	**48.8**	**50.1**	**84.8**	**60.2**
VideoPrism-g (Zhao et al., 2024)	288	16	-	76.4	-	-	-	50.3	52.7	71.0	39.7	-	-
InternVideo2-6B (Wang et al., 2024b)	224	8	-	-	-	-	62.0	56.5	63.2	53.7	55.9	83.1	59.3

3.2.4 MULTIMODAL UNDERSTANDING

We compare vision towers using the same two-stage LLaVA-NeXT recipe, Qwen2.5-7B-Instruct, and 19 evaluation tasks. Only the projector is aligned in Stage-1, and all parameters are updated in Stage-2. Table 3 reports overall and task-group means. Under Square inputs, H exceeds SigLIP 2-SO by 0.53 overall, largely from HallusionBench, while its OCR/Chart/Document average decreases. G exceeds SigLIP 2-g by 0.95, improving all three groups. Square comparisons match patch tokens, with COKE supplying 12 additional register/reference tokens to the language model. Tiling improves the overall score by 4.08 for H and 1.21 for G, primarily through OCR/Chart/Document tasks. Native adds 0.18 and 0.46 over the corresponding tiled configurations at matched total visual-token counts. Different checkpoints and attention layouts prevent attributing these differences solely to input processing or assuming equal encoder FLOPs. Table 3 reports all task scores.

Table 3: **Multimodal understanding under LLaVA-NeXT.** Scores are normalized to 0 to 100. **Overall** and **Average** give unweighted means over 19 tasks and within each group. Blockwise column maxima are bold. Square matches patch counts, adding 12 COKE REG/REF tokens. Native matches its COKE tiled counterpart's total visual tokens per image. Inputs and metrics follow §C.5.1.

				General Understanding									OCR / Chart / Document								Reasoning				
Vision Encoder	**Res./Patch**	**Input Strategy**	**Overall**	**Average**	RealWorldQA	SEEDBench	MME	MMBench	VizWiz	RefCOCO	POPE	HallusionBench	**Average**	AI2D	ChartQA	DocVQA	InfoVQA	TextVQA	OCRBench	OCRBench v2	**Average**	GQA	MMMU	MMStar	MathVista
Patch tokens 576																									
MetaCLIP 2-H	336/14	Square	52.03	58.16	59.3	71.1	68.8	66.5	59.3	27.7	88.3	24.3	49.01	71.9	59.0	53.9	32.1	52.5	44.8	28.9	45.05	60.4	45.7	47.1	27.0
PE_{core}-L	336/14	Square	53.22	57.77	58.6	73.3	**71.7**	65.9	58.4	26.5	**88.9**	18.9	52.29	73.4	64.0	60.0	32.6	56.2	49.8	30.0	45.75	61.8	46.8	47.1	27.3
SigLIP-SO	384/16	Square	53.73	58.00	62.2	73.3	69.9	67.5	58.6	25.6	88.0	18.9	53.14	73.9	65.8	**62.4**	34.9	58.1	46.8	30.1	46.20	61.6	46.9	48.9	27.4
SigLIP 2-SO	384/16	Square	54.66	58.70	61.6	73.6	70.6	**68.8**	60.8	27.3	88.0	18.9	**54.49**	**75.7**	**66.7**	61.4	**35.4**	**60.7**	**50.5**	31.0	46.90	62.0	**48.0**	50.2	27.4
COKE-ViT-H	384/16	Square	**55.19**	**60.38**	**62.7**	**74.3**	70.1	68.7	**61.8**	**28.2**	88.5	**28.7**	53.80	75.4	66.6	60.8	34.2	58.5	49.8	**31.3**	**47.25**	**62.1**	47.8	**51.3**	**27.8**
High-resolution inputs based on 384/16																									
SigLIP 2-SO	384/16	Split-Tile	58.90	60.88	62.5	74.5	71.0	68.4	61.3	28.6	89.5	**31.2**	63.06	76.9	76.2	81.4	**46.4**	**70.1**	56.4	34.0	47.67	62.8	47.8	51.1	29.0
COKE-ViT-H	384/16	Split-Tile	59.27	**61.61**	**67.2**	74.9	70.8	**69.5**	**63.2**	28.1	**89.8**	29.4	62.87	75.0	75.4	80.9	44.6	68.7	60.6	34.9	**48.27**	**63.1**	**48.3**	**52.7**	29.0
COKE-ViT-H	Native	Native	**59.45**	61.14	64.6	**75.2**	**71.9**	68.4	61.0	**29.6**	89.4	29.0	**64.09**	**77.1**	**77.0**	**83.8**	44.4	69.8	**61.2**	**35.3**	47.98	62.5	47.7	52.4	**29.3**
Patch tokens 1024																									
PE_{core}-G	448/14	Square	56.79	60.80	63.0	74.0	**71.8**	67.6	60.7	**30.1**	89.6	29.6	57.74	75.6	68.8	70.1	38.3	62.7	56.4	32.3	47.12	62.3	**48.0**	50.8	27.4
SigLIP 2-g	512/16	Square	57.69	60.86	63.5	**75.2**	71.5	**70.0**	**62.9**	29.1	89.1	25.6	59.63	76.7	71.8	71.2	38.8	66.3	59.4	33.2	47.98	**63.5**	47.2	**53.4**	27.8
COKE-ViT-G	512/16	Square	**58.64**	**61.71**	**64.1**	**75.2**	71.5	69.5	62.1	28.5	**89.7**	**33.1**	**61.11**	**77.2**	**75.0**	**74.2**	**40.3**	**66.7**	**60.1**	**34.3**	**48.17**	62.9	46.7	53.2	**29.9**
High-resolution inputs based on 512/16																									
SigLIP 2-g	512/16	Split-Tile	59.67	61.44	67.7	**75.5**	**72.3**	**70.0**	57.2	28.5	89.8	**30.5**	64.26	76.2	75.6	84.1	46.0	**71.9**	61.5	34.5	48.12	**63.3**	46.2	**54.6**	28.4
COKE-ViT-G	512/16	Split-Tile	59.85	60.98	**68.8**	75.3	70.8	69.8	62.3	27.1	**90.0**	23.7	65.16	75.9	74.7	84.9	49.2	71.5	**63.7**	**36.2**	48.30	62.7	48.3	52.6	29.6
COKE-ViT-G	Native	Native	**60.31**	**61.66**	66.0	74.7	72.1	68.8	**62.4**	**29.6**	89.5	30.2	**65.47**	**76.3**	**77.8**	**86.6**	**49.6**	71.1	61.5	35.4	**48.55**	62.8	**48.6**	53.0	**29.8**

3.3 Ablation Studies

Image-side ablations use COKE-ViT-H-256 with 1.6B samples each, except the 400M distillation-weight sweep. A separate convergence study uses SO400M-384. Video strategy, hyperparameter, and vocabulary experiments use stated budgets. Token counts and Snapshot-Replay parameters were chosen using aggregate scores on reported benchmarks with no separate validation set and are not independently validated optima. Appendix C.6 gives full settings and additional results.

3.3.1 Evaluation of Visual Reference Pretrain

Table 4(a) evaluates token branches. Registers improve the overall score from 74.73 to 75.42, reference distillation alone reaches 78.54, and their combination reaches 79.43. With teacher, loss weight, and sample budget matched, pooled-patch distillation improves classification but reduces retrieval relative to no distillation in Table 4(b). Reference-token supervision exceeds pooled-patch supervision by 2.80 classification and 5.20 retrieval points. This comparison also adds eight reference tokens, so token addition and output choice are not isolated. Table 4(c) examines distillation weights at a separate 400M-sample budget. Reducing the pooled-patch weight from 4000 to 1000 improves retrieval but lowers classification. Reference-token distillation exceeds the best pooled-patch classification and retrieval means across tested weights by 1.38 and 2.33 points, respectively. Thus, reducing the pooled-patch weight does not close the observed gap within this sweep. Full results and early sample efficiency appear in Appendices D.5 and D.2, respectively.

Table 4: **VRP design comparisons.** (a) Token branches and (b) distillation pathways use 1.6B samples per run. (c) Distillation weights are compared at a separate 400M budget. Overall averages ten image metrics, while CLS and RET average six classification and four retrieval scores. Panels (b) and (c) use four registers, with eight reference tokens only in reference configurations. Panel (b) matches teacher and loss weight. Full results appear in Tables 13, 16, and 21.

(a) Token branches

REG	*REF*	Overall
✗	✗	74.73
✓	✗	75.42
✗	✓	78.54
✓	✓	79.43

(b) Distillation pathways

Pathway	CLS	RET
None	76.48	73.83
Pooled patch	79.55	69.85
Reference	82.35	75.05

(c) Distillation weight

Pathway	λ	CLS	RET
Pooled patch	400	67.65	70.25
Pooled patch	1000	70.33	70.30
Pooled patch	4000	72.77	68.33
Reference	4000	**74.15**	**72.63**

3.3.2 Evaluation of Snapshot-Replay

Strategies in Table 5(a) share image-pretrained COKE-ViT-H-256 initialization and 640M video-text samples without image-text replay. WiSE-FT interpolates initialization and the continuous fine-tuning endpoint ($\beta = 0.5$). Snapshot-Replay uses final fused weights. Continuous fine-tuning raises the video mean (56.23 to 61.20) but lowers the image mean (81.55 to 81.21). Lookahead remains below the initial checkpoint in image performance. Snapshot-Replay averages 83.02 on images and 62.23 on videos, exceeding continuous fine-tuning by 1.81 and 1.03 and WiSE-FT sweep maxima by 0.23 and 0.46 points, respectively (Appendix D.4). These gains support image retention and video adaptation in this setting. Panels (b,c) vary burst length K and snapshot weight α with 128M samples. At $\alpha = 0.75$, $K = 10$ improves image but lowers video performance versus $K = 1$. $K = 100$ underperforms $K = 10$ on both. At $K = 10$, image scores peak at $\alpha = 0.75$ and video scores at 0.01. $\alpha = 0.99$ lowers both versus 0.75. The chosen $K = 10$ and $\alpha = 0.75$ favor image performance with competitive video results, without maximizing both metrics. Appendix C.6.8 gives the setup.

Table 5: **Snapshot-Replay ablations.** Runs share image-pretrained COKE-ViT-H-256 initialization. (a) Strategies use 640M video-text samples. (b) Burst length K and (c) snapshot weight α vary over 128M samples with $\alpha = 0.75$ and $K = 10$ fixed, respectively. Image/Video uniformly average ten image/nine video metrics, counting retrieval directions separately. Compare within panels.

(a) Fine-tuning strategies

Strategy	Image	Video
Image-Pretrained	81.55	56.23
Video-only Finetune	81.21	61.20
Lookahead	81.51	61.31
WiSE-FT ($\beta = 0.5$)	82.79	61.01
Snapshot-Replay (Ours)	**83.02**	**62.23**

(b) Burst length

K	Image	Video
1	82.55	**61.76**
10	**82.90**	61.24
100	82.51	59.71

(c) Snapshot weight

α	Image	Video
0.99	81.81	57.59
0.75	**82.90**	61.24
0.01	82.16	**61.61**

3.3.3 Effectiveness of Vocabulary-Based Data Selection

Table 6 evaluates COKE-ViT-H-256 with a randomly initialized visual backbone, a frozen pretrained SigLIP 2 text tower, and no VRP. All runs use 1B samples with identical optimization and evaluation. Vocabulary-based selections share matching and sampling. On DataComp-1B, MetaCLIP exceeds random selection by 1.15 overall. On the public pool, COKE-K exceeds MetaCLIP by 0.73 in classification, 0.89 in retrieval, and 0.80 overall, with pool size, sampling cap, and schedule fixed. COKE-V leads classification and overall performance, and COKE-K leads retrieval. COKE-V exceeds MetaCLIP by 2.48 overall but uses a smaller, repeatedly sampled pool, so its gain does not isolate vocabulary choice from data reuse.

Table 6: **Vocabulary-based selection on 1B samples.** COKE-ViT-H-256 freezes the text tower without VRP. CLS averages six classification scores, RET four retrieval scores, and Image all ten. Parentheses give `max_match`. Four-epoch runs draw 0.25B random pairs/epoch from chosen pools.

Data Source	Sample Strategy	Selected Pool	Epochs	CLS	RET	Image
DataComp-1B	Random	0.25B	4	70.33	69.58	70.03
DataComp-1B	MetaCLIP (10K)	0.26B	4	71.45	70.78	71.18
Public pool	MetaCLIP (40K)	1B	1	70.31	71.49	70.78
Public pool	COKE-K (40K)	1B	1	71.04	72.38	71.58
Public pool	COKE-V (40K)	0.28B	4	74.03	72.11	73.26

4 Conclusion

COKE unifies image-video encoding using public corpora and released model priors. Visual Reference Pretrain improves early sample efficiency, with reference-token supervision outperforming tested pooled-patch distillation in classification and retrieval. Snapshot-Replay resumes optimization after periodic interpolation toward a fixed image checkpoint, improving image and video performance over continuous video-only fine-tuning under matched initialization and data. Fixed-budget ablations support COKE-K and COKE-V balancing, combined with quality filtering and deduplication to construct COKE-1.7B. COKE supports fixed/native resolutions with competitive zero-shot classification, retrieval, attentive probing, and multimodal transfer. Video performance trails PE-G in some settings, and native inputs yield modest, task-dependent gains over tiling.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure, namely interpretation of results and assistance with translation. Specifically, we used the released Seed-X model to translate multilingual captions into English during data curation (see Appendix B.3), and we used generative AI tools to help interpret and phrase the reported results. We have not used generative AI tools for the remaining tasks with required disclosure, such as generating synthetic data, formulating or proving mathematical claims, proposing hypotheses, or designing the core methods and experiments. Additionally, we used generative AI tools for tasks with recommended disclosure, including editing the paper to improve readability, suggesting the paper structure, drafting parts of the paper, and creating or modifying scientific figures and tables. All AI-assisted text was reviewed by all authors, and AI-assisted figures and tables were verified against the experimental logs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we report the model architecture in Appendix B.1, the training schedule and hyperparameters in Appendix B.2 and Table 9, and the data curation pipeline in Appendix B.3. Evaluation protocols for zero-shot image and video classification and retrieval, frozen-encoder probing, and multimodal understanding are described in Appendix C and its subsections. Our training relies on public image-text corpora (DataComp-XL, LAION-5B, COYO-700M) and public video-text corpora (Panda-70M, WebVid-10M, InternVid-10M-FLT), together with released auxiliary models (Seed-X, SigLIP 2, and DINOv3), as summarized in Appendix D.6. All experiments are implemented in PyTorch with DeepSpeed, FlashAttention-2, and activation checkpointing. We plan to release model weights, the COKE-K and COKE-V vocabularies, and evaluation code.

ETHICS STATEMENT

Our work trains large vision-language encoders on publicly crawled image-text and video-text corpora. These corpora may contain personally identifiable information, harmful content, and demographic or societal biases; our quality filtering, concept balancing, and deduplication do not specifically remove such content, so we do not claim that the released models are bias-mitigated or safe for high-stakes deployment. We recommend additional filtering and model auditing before any downstream use. The experiments involve no human subjects and no private data collection. Training consumes substantial computational resources (approximately 10K and 13K device-days on H800 GPUs for COKE-ViT-H and COKE-ViT-G, respectively), with associated energy and carbon costs. We have read and will adhere to the ICLR Code of Ethics.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, 32, 2019.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 190–200, 2011.

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? *arXiv preprint arXiv:2403.20330*, 2024a.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan Wang, and Yonghui Wu. Seed-x: Building strong multilingual translation llm with 7b parameters. *arXiv preprint arXiv:2507.13618*, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen-tau Yih, Shang-Wen Li, and Hu Xu. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16588.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023a.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36: 2252–2274, 2023b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

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

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

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. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Drew A Hudson and Christopher D Manning. 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.

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. Version 0.1. `https://doi.org/10.5281/zenodo.5143773`, 2021. DOI: 10.5281/zenodo.5143773.

Filip Ilievski, Pedro Szekely, and Bin Zhang. Cskg: The commonsense knowledge graph. In *European Semantic Web Conference (ESWC)*, pp. 680–696. Springer, 2021.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, pp. 4651–4664, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023c.

Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1% zero-shot imagenet accuracy within $10000 budget; an extra $4000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*, 2023d.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834. PMLR, 2018. URL `https://proceedings.mlr.press/v80/li18a.html`.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23390–23400, 2023e.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023f.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pp. 1697–1706, 2022.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Christian M. Meyer and Iryna Gurevych. Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot (eds.), *Electronic Lexicography*, pp. 259–291. Oxford University Press, 2012. doi: 10.1093/acprof:oso/9780199654864.003.0013.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 2022.

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, 2024.

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.

Limeng Qiao, Yiyang Gan, Bairui Wang, Jie Qin, Shuang Xu, Siqi Yang, and Lin Ma. Vitrix-univitar: Unified vision transformer with native resolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model – reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, pp. 3505–3506, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Michaël Ramamonjisoa, Tatjana Chavdarova, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.

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

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, volume 31, pp. 4444–4451, 2017.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023a.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023b.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL `https://proceedings.mlr.press/v139/touvron21a.html`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024b.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, pp. 23965–23998, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022b.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, 2024. RealWorldQA benchmark.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pp. 5288–5296, 2016.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pp. 69–85, 2016.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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.

Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. Tvtsv2: Learning out-of-the-box spatiotemporal visual representations at scale. *arXiv preprint arXiv:2305.14173*, 2023.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.

A Related Work

A.1 Contrastive Vision-Language Foundation Models

Objectives, Scale, and Native-Resolution. CLIP (Radford et al., 2021) established image-text contrastive learning as the default recipe for open-vocabulary visual representations, and later work advanced it along several largely independent directions. The objective was refined by the sigmoid formulation of SigLIP (Zhai et al., 2023), which replaces the global softmax with pairwise sigmoid supervision. EVA-CLIP (Sun et al., 2023b) and OpenCLIP (Ilharco et al., 2021; Cherti et al., 2023) study improved training recipes and scaling for contrastive encoders. SigLIP 2 (Tschannen et al., 2025) and Perception Encoder (Bolya et al., 2025) further develop comprehensive pretraining pipelines. CLIPA-v2 (Li et al., 2023d) focuses on reducing training cost through token reduction and efficient fine-tuning. NaViT (Dehghani et al., 2023b) introduces sequence packing to accommodate images with varying resolutions and aspect ratios. Qwen2-VL (Wang et al., 2024a) and the NaFlex variant of SigLIP 2 (Tschannen et al., 2025) further develop variable-resolution visual encoding. UniViTAR (Qiao et al., 2025) combines fixed-to-native resolution training, early visual-teacher distillation, and image-video alternating updates within a unified encoder. Building on this foundation, COKE studies reference-token supervision for the classification–retrieval trade-off, replaces alternating updates with video-only branches using fixed-checkpoint interpolation, and evaluates knowledge-based vocabularies under fixed training budgets.

Unified Encoders for Image and Video. Extending image-text encoders to video commonly involves temporal modules or video-specific training. CLIP-ViP (Xue et al., 2022) adapts CLIP through video-text post-training, while TVTSv2 (Zeng et al., 2023) studies transferable spatiotemporal representations. VideoCoCa (Yan et al., 2022) adapts CoCa's attentional pooling to frame embeddings. UMT (Li et al., 2023c) uses an unmasked image teacher to supervise masked video modeling. InternVideo2 (Wang et al., 2024b) combines masked video modeling with multimodal alignment, while VideoPrism (Zhao et al., 2024) follows video-text pretraining with masked video modeling. Omnivore (Girdhar et al., 2022) demonstrates shared-parameter visual modeling across images, videos, and single-view 3D inputs using spatiotemporal patches. Perception Encoder (Bolya et al., 2025) adapts an image encoder to video-text learning by averaging frame embeddings. COKE independently encodes consecutive frame pairs as spatiotemporal cubes and averages their features over time. Frame duplication allows images to use the same patchifier and encoder (§2.1).

Vision Encoders as Multimodal LLM Front Ends. Vision encoders provide visual representations for multimodal language models. LLaVA (Liu et al., 2024a) connects a pretrained CLIP encoder to a language model through a lightweight projector, and subsequent systems such as InternVL (Chen et al., 2024c) and Qwen2-VL (Wang et al., 2024a) further develop this paradigm. Cambrian-1 (Tong et al., 2024) systematically evaluates visual representations through multimodal language modeling, while MM1 (McKinzie et al., 2024) studies the effects of the image encoder, resolution, and visual token count. These studies motivate evaluating COKE under a shared multimodal training recipe and explicitly reporting the visual token budgets of the compared configurations (§3.2.4).

A.2 Pretrained Knowledge Transfer and Retention

Visual Supervision and Token-Based Distillation. Non-patch tokens have several roles in vision transformers. ViT uses a *[CLS]* token to aggregate global features (Dosovitskiy et al., 2020), while register tokens provide additional computational capacity within the attention sequence (Darcet et al., 2024). DeiT introduces a teacher-supervised distillation token interacting with other tokens through attention (Touvron et al., 2021). Perceiver latent arrays (Jaegle et al., 2021) and BLIP-2 queries (Li et al., 2023b) provide related interfaces for feature aggregation. SLIP (Mu et al., 2022) combines self-supervised learning with image-text contrastive pretraining. SILC (Naeem et al., 2024) introduces local-to-global self-distillation with an EMA teacher, while AM-RADIO (Ranzinger et al., 2024) distills multiple pretrained vision models into a single student. DeiT distills class predictions, whereas COKE supervises reference outputs with frozen-teacher features during early image-text pretraining and aligns pooled patch outputs with text in a shared backbone. We study the classification–retrieval trade-off of this output choice (§3.3.1).

Contrastive Alignment with Pretrained Encoders. LiT (Zhai et al., 2022) studies contrastive alignment with pretrained encoders and finds that freezing the image encoder while training the text

encoder is effective for zero-shot transfer. COKE instead freezes a pretrained text tower during VRP while training the visual encoder. We evaluate text-tower initialization and update policies in §C.6.6.

Weight-Space Retention of Prior Capability. Weight averaging and interpolation are established approaches to improving transfer and retaining pretrained capability. SWA averages optimization iterates (Izmailov et al., 2018), model soups combine fine-tuned checkpoints (Wortsman et al., 2022a), and WiSE-FT interpolates pretrained and fine-tuned weights after training (Wortsman et al., 2022b). Lookahead periodically interpolates fast weights with a moving slow state and resumes training from the result (Zhang et al., 2019). EMA maintains running weight averages, also used for teacher supervision in Mean Teacher (Tarvainen & Valpola, 2017). L2-SP penalizes deviations from pretrained weights in the training objective (Li et al., 2018). Snapshot-Replay instead alternates video-only optimization with periodic interpolation toward a fixed image-pretrained checkpoint. Its fixed target differs from Lookahead's moving state, and resuming video optimization distinguishes it from WiSE-FT and our evaluation-only EMA baseline (§2.3).

A.3 Data Curation for Image-Text Pretraining

Public pools such as LAION-5B (Schuhmann et al., 2022), COYO-700M (Byeon et al., 2022), and DataComp (Gadre et al., 2023) provide large-scale image-text data. DataComp also offers a framework to compare data curation methods under standardized model and training settings. DFN (Fang et al., 2023) studies model-based filtering, while MetaCLIP (Xu et al., 2023) balances samples against metadata from WordNet and Wikipedia. MetaCLIP 2 extends metadata-based curation to multilingual data (Chuang et al., 2025). COKE follows MetaCLIP's concept-matching and balancing approach, using COKE-K derived from CSKG and COKE-V derived from ImageNet-21K categories in a pipeline with quality filtering and deduplication (§2.4). At a fixed cumulative sample budget, we compare COKE-K and COKE-V with MetaCLIP on the same candidate pool using shared matching rules and sampling caps (§3.3.3).

B Supplementary Method Details

B.1 Model Architecture Details

Table 7 summarizes the architecture settings of COKE-ViT-H and COKE-ViT-G. The variants scale hidden width, depth, attention head dimension, and SwiGLU hidden dimension while retaining the shared settings listed on the right. Both omit the *[CLS]* token and form image embeddings from pooled patch features. Four register tokens and eight reference tokens are used at both scales.

Table 7: **Architecture configuration of the COKE-ViT family.** The COKE-ViT-H and COKE-ViT-G variants differ only in the scale-dependent quantities on the left, while every other design choice is shared on the right.

Scale-dependent	COKE-ViT-H	COKE-ViT-G
Num. Parameters	0.6B	1.9B
Hidden size d	1280	1792
Num. layers L	32	48
Head dim.	80	112
SwiGLU hidden dim.	3456	5120

Shared across scales	
Num. attention heads	16
Spatial patch size p_s	16
Norm / FFN	RMSNorm / SwiGLU
Positional encoding	RoPE-2D
[CLS] / *[REG]* / ***[REF]***	0 / 4 / **8**

Patch embedding and transformer blocks. For T sampled frames of size $H \times W$, the encoder forms $T' = T/p_t$ temporal cubes with $p_t = 2$ frames each. Non-overlapping patches of size (p_t, p_s, p_s) are linearly embedded into d dimensions, yielding $H'W'$ patch tokens per cube, where $H' = H/p_s$ and $W' = W/p_s$. The backbone comprises L pre-norm transformer layers. Each layer combines masked multi-head self-attention with QK-normalization (Dehghani et al., 2023a) and RoPE-2D (Su et al., 2024) with a SwiGLU feed-forward block, using RMSNorm (Touvron et al., 2023) for normalization. RoPE-2D applies only to patch tokens. Register and reference tokens receive no rotary positional encoding and interact with patch tokens through self-attention.

B.2 Training Configuration

Fig. 3 summarizes the image-training sequence and independent video branches described in §2.1. Only image-text checkpoints initialize subsequent resolution levels. At each level, a separate Snapshot-Replay branch adapts the corresponding image-text checkpoint using video-text pairs, producing weights shared by image and video tasks.

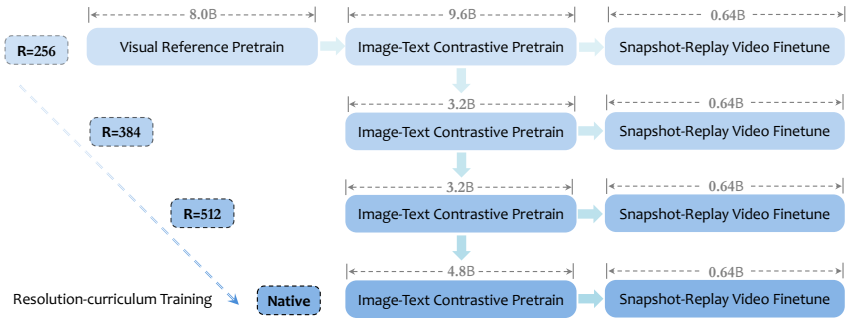


Figure 3: **Training strategy of COKE-CLIP.** Image-text training progresses from R=256 through 384 and 512 to native resolution, with Visual Reference Pretrain applied at the beginning of the first level. At each level, the image-text checkpoint also initializes an independent video-text Snapshot-Replay branch, yielding an encoder that supports both images and videos. Only the image-text checkpoints initialize subsequent resolution levels.

Resolution curriculum. Following fixed-to-native training approaches (Dehghani et al., 2023b; Wang et al., 2024a; Qiao et al., 2025), the four levels use R=256, 384, 512, and native resolution, with total image and video sample budgets of 18.24B, 3.84B, 3.84B, and 5.44B. Concentrating most image-text samples at the lowest resolution is intended to control computation before adaptation to higher spatial detail and variable aspect ratios. Patch-token count grows quadratically with image resolution, and full self-attention includes a term quadratic in token count. The curriculum is evaluated within the complete recipe rather than through an isolated schedule ablation.

RoPE-2D uses coordinates rather than a learned positional table, and packing supports variable spatial grids. Image checkpoints therefore initialize the next level without an architectural change. Native-resolution inputs are limited to 16384 patch tokens per sample before patch dropout and resized when necessary to satisfy this budget. Table 8 reports the sample allocation, and Table 9 provides the per-stage hyperparameters.

B.3 Data Curation Details

Fig. 4 provides the complete pipeline underlying the concept-based selection described in §2.4. The following details specify the source pool, vocabulary construction, sampling thresholds, and intermediate corpus sizes.

Data pool. We aggregate DataComp-XL (Gadre et al., 2023), LAION-5B (Schuhmann et al., 2022), and COYO-700M (Byeon et al., 2022) into a raw pool of approximately 18.7B image-text pairs. Re-crawling and image-validity screening leave approximately 15B pairs. Captions from the LAION-2B-multi and LAION-1B-nolang subsets are translated into English using the released Seed-X model (Cheng et al., 2025). Translation allows captions from these multilingual sources to be matched against the same English concept vocabularies. The resulting approximately 15B-pair pool forms the input to the score-based filtering and deduplication stage described below.

Common-sense knowledge vocabulary. COKE-Vocab consists of two inventories derived from public knowledge resources. COKE-K is built from CSKG (Ilievski et al., 2021), which consolidates resources including ConceptNet (Speer et al., 2017), Wiktionary (Meyer & Gurevych, 2012), and WordNet (Miller, 1995). We retain uni- and bi-gram entries and discard phrases with three or more words, yielding 1,099,017 entries. COKE-V contains approximately 21K entries from the ImageNet-21K (Deng et al., 2009) and provides an explicitly visual category inventory. Both inventories serve as textual entries for caption matching and subsequent sampling. In particular, COKE-V uses ImageNet

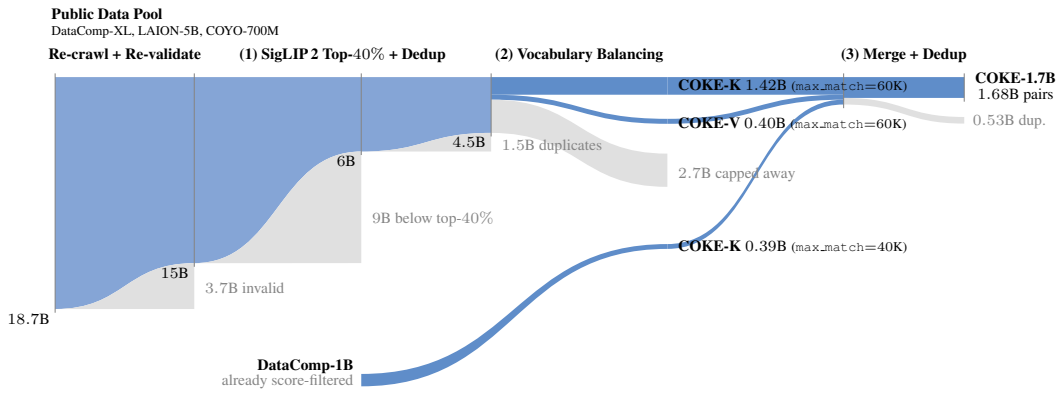


Figure 4: **Overview of the data curation pipeline.** Band thickness shows the number of image-text pairs, and grey bands signify discarded samples. Filtering and deduplication reduce the 18.7B public pool to a 4.5B candidate pool, from which COKE-K and COKE-V produce separate selections. A third selection is sampled from DataComp-1B using COKE-K. Global deduplication reduces the potentially overlapping selections from 2.21B to 1.68B pairs, forming **COKE-1.7B**.

category names to retrieve matching image-text pairs from the candidate pool. Vocabulary size describes the available matching entries, while actual concept coverage depends on their occurrence in captions and retention during curation.

Curation pipeline. Using the data pool and COKE-Vocab, we construct the training corpus in three stages (Fig. 4). Deduplication uses URLs and image-content hashes. *(1) Score-based filtering and deduplication.* We score the approximately 15B valid image-text pairs using SigLIP 2 (Tschannen et al., 2025) and retain the top 40% ranked by image-text similarity, yielding approximately 6B pairs. Deduplication removes a further 25%, leaving approximately 4.5B candidate pairs. This stage filters low-scoring and duplicate pairs. The next stage balances concept frequencies. *(2) Vocabulary-based balancing.* We follow MetaCLIP (Xu et al., 2023) for text preprocessing and vocabulary matching. For each COKE-Vocab entry, we retain all matching pairs if their count does not exceed `max_match`, and otherwise uniformly sample `max_match` pairs without replacement. Since captions may match multiple entries, selections are merged and deduplicated per sampling pass. The cap applies to each concept-specific selection before merging, so a concept's frequency in the merged selection may exceed `max_match`. This procedure complements similarity-based filtering by limiting frequent concepts' contributions during sampling. We run three passes. Two use the 4.5B pool, balanced separately against COKE-K and COKE-V, both with `max_match`=60K, and yield 1.42B and 0.40B pairs respectively. The third instead samples the original DataComp-1B dataset directly against COKE-K with `max_match`=40K, yielding 0.39B pairs, and bypasses (1) because DataComp-1B is already score-filtered. *(3) Merging and global deduplication.* Although each sampling pass deduplicates its selection, the three selections may still share image-text pairs. Before cross-selection deduplication, they total 2.21B. Removing overlaps reduces this total by approximately 24%, yielding the **COKE-1.7B** corpus of 1.68B image-text pairs for training the main models. Deduplication covers only the training corpus, without additional benchmark-specific decontamination. Each training pass draws 1.6B pairs from it, so the 28.8B image-text samples correspond to 18 such passes.

C Supplementary Experimental Details and Results

This section provides the complete settings and results underlying the main-text summaries. We evaluate COKE along four axes. Zero-shot image classification and retrieval (§C.2) assess the joint image-text embedding, whereas frozen-encoder probing (§C.4) isolates the visual representation alone. Zero-shot video benchmarks (§C.3) test temporal transfer, and multimodal understanding (§C.5) assesses MLLM vision-tower utility. Ablations (§C.6) assess the contributions of individual components and training configurations.

C.1 Training Recipe

C.1.1 Data Details

Image-text data. Image-text training of the main models uses the COKE-1.7B corpus curated in §2.4, which is assembled entirely from publicly released datasets. Each epoch draws 1.6B pairs, so the 28.8B image-text samples of the schedule amount to roughly 18 passes over the corpus.

Video-text data. We collect three public video-text datasets, namely Panda-70M (Chen et al., 2024b), WebVid-10M (Bain et al., 2021), and InternVid-10M-FLT (Wang et al., 2023), and retain their captions exactly as released. Since these corpora distribute video URLs rather than videos, discarding unreachable or corrupt clips leaves roughly 64M pairs, with no further filtering. The contribution of Snapshot-Replay is evaluated through controlled strategy ablations using the same video data and initialization. Each sample pairs one clip with its caption. We uniformly sample 16 frames across each video clip and group consecutive frames in the sampled sequence into 8 temporal cubes, each containing p_t=2 frames. Frames follow the same resolution regime as images. At the three fixed levels, each frame is resized to the corresponding resolution. At the native level, frame dimensions are adjusted so that the entire clip fits within a budget of 16384 patch tokens before patch dropout.

Budget allocation. The schedule of §2.1 consumes 28.8B image-text and 2.56B video-text samples, or 31.4B in total. Following the equal-share design of §2.3, the video budget is divided evenly across the four resolution levels at 640M samples each, which amounts to 10 passes over the 64M video-text pairs per level. Subtracting this share from the per-level totals in Appendix B.2 leaves 17.6B, 3.2B, 3.2B, and 4.8B image-text samples at R=256, 384, 512, and native resolution respectively, where the first 8B at R=256 constitute the Visual Reference Pretrain stage of §2.2. Consequently 61% of the image-text budget is spent at the cheapest level, as Table 8 details.

Table 8: **Sample budget allocation.** Image-text samples are drawn from COKE-1.7B and video-text samples from the 64M public collection. Levels receive equal video-text shares, with most image-text samples allocated to the lowest resolution.

Level	R=256	R=384	R=512	Native	**Total**
Image-Text	17.6B	3.20B	3.20B	4.80B	28.8B
Video-Text	0.64B	0.64B	0.64B	0.64B	2.56B
Total	18.24B	3.84B	3.84B	5.44B	31.36B

C.1.2 Hyperparameter Details

Shared configuration. All stages are optimized with AdamW (β_1=0.9, β_2=0.95, weight decay 0.1) under a cosine-decayed learning rate with linear warmup and gradient clipping at 1.0. The learnable logit scale τ and bias b in (1) are initialized as in SigLIP 2 (Tschannen et al., 2025). The maximum text length is 64 tokens. Training runs under DeepSpeed ZeRO Stage 1 (Rasley et al., 2020) with activation checkpointing (Chen et al., 2016) and FlashAttention-2 (Dao, 2024). At the three fixed levels, images are augmented with `RandomResizedCrop` at scale=$(0.9, 1.0)$ and bicubic interpolation. At the native level, images exceeding the budget of 16384 patch tokens are resized to fit within it before patch dropout. Spatial dimensions are adjusted to multiples of the patch size 16, approximately preserving the original aspect ratio. All models are trained on 256 H800 GPUs, and the full schedule of Table 8 takes $10K$ and $13K$ device-days for COKE-ViT-H and COKE-ViT-G respectively.

Per-stage configuration. Table 9 summarizes settings across training stages. Patch dropout (Li et al., 2023e) removes a fraction of patch tokens during training, while inference uses the full grid. The dropout rate is 0.5 for image-text training and 0.75 for video-text fine-tuning. We use a default distillation weight of $\lambda = 4000$ in (1), unless otherwise specified. Snapshot-Replay uses K=10 and α=0.75 across four resolutions.

Table 9: **Per-stage training hyperparameters.** Columns summarize the stages in §2.1. Learning rates are reported as peak → final under cosine decay. During VRP, randomly initialized COKE-ViT-H and COKE-ViT-G use peak vision learning rates of 1×10^{-3} and 8×10^{-4}, respectively. Output heads use the vision-branch learning rate. One epoch denotes 1.6B image-text samples from COKE-1.7B or 64M video-text samples. Shared settings are specified in the text.

Config	Visual Reference Pretrain	Image-Text Contrastive Training		Video-Text Snapshot-Replay Finetune
	R=256	R=256	R=384/512/Native	R=256/384/512/Native
Vision Tower LR	1e-3 / 8e-4 → 1e-6	1e-5 → 0	1e-5 → 0	1e-6 → 0
Text Tower LR	Frozen	1e-6 → 0	1e-6 → 0	1e-7 → 0
Warmup Steps	10000	10000	2000	2000
Global Batch Size	32768	32768	32768	8192
Patch Dropout	0.5	0.5	0.5	0.75
Epochs	5	6	2 / 2 / 3	10
λ	4000	-	-	-
(K, α)	-	-	-	(10, 0.75)

C.2 Results on Zero-shot Image Classification & Retrieval

C.2.1 Evaluation Setup

We evaluate zero-shot classification on six benchmarks, namely ImageNet-1K (Deng et al., 2009), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), and ObjectNet (Barbu et al., 2019). Zero-shot retrieval is evaluated on Flickr30K (Young et al., 2014) and MSCOCO (Lin et al., 2014). Classification is scored by top-1 accuracy and retrieval by Recall@1. For models evaluated in this work, we adopt the protocol of PE_{core} (Bolya et al., 2025) in full. Class names are therefore expanded with the task-specific ImageNet prompts of InternVL (Chen et al., 2024c), and the resulting text embeddings are averaged into a single classifier weight per class, whereas retrieval uses the original captions without any prompt. For models evaluated in this work, we follow (Sun et al., 2023a) in evaluating each dataset under both center crop and squash resize, reporting the higher of the two scores. Retrieval scores follow the reweighting implementation of PE_{core} (Bolya et al., 2025). Starting from the cosine-similarity matrix $S \in \mathbb{R}^{N_{\text{text}} \times N_{\text{media}}}$, the text-to-media and media-to-text directions are reweighted separately as $S \odot \text{softmax}_{\text{dim}=0}(S)$ and $S^\top \odot \text{softmax}_{\text{dim}=0}(S^\top)$. Thus, softmax operates over the query dimension in each direction, and retrieval ranks candidates along the gallery dimension. We retain the default `reweight_scale`$= 1.0$ and apply this procedure consistently to all models we re-evaluate. Under this pipeline we re-evaluate SigLIP 2 (Tschannen et al., 2025) and PE_{core} (Bolya et al., 2025), marked with † in Table 10. Other baseline results are taken from UniViTAR (Qiao et al., 2025) where available and otherwise from the original publications.

C.2.2 Results Comparison and Analysis

Zero-shot classification and retrieval. Table 10 compares models at fixed and native resolutions. At 384, COKE-CLIP-H achieves classification and retrieval averages of 85.9 and 80.4, compared with 84.6 and 77.4 for SigLIP 2-SO. At 512, COKE-CLIP-G obtains 86.8 and 79.9, compared with 86.6 and 78.9 for PE_{core}-G. For H, the largest classification gain occurs on ImageNet-A (92.1 vs. 87.0), while all four retrieval metrics improve. G nearly matches PE_{core}-G in classification, with a clearer advantage in retrieval (79.9 vs. 78.9). These results show competitive classification and improved retrieval across directions.

Resolution and model scale. Classification improves across the fixed-resolution checkpoints, with gains of 0.9 and 0.8 points from 256 to 384 for H and G, followed by 0.2 points for both at 512. Native resolution yields the highest retrieval averages (80.8 and 80.5), while classification is 0.7 points below the corresponding 512 checkpoints at both scales. Across all four resolution settings, G scores higher than H on classification but lower on retrieval. These patterns indicate task-dependent benefits from model scale and resolution. Since the checkpoints follow progressive training, their differences also include the effects of additional training.

Training resources. COKE-1.7B contains fewer unique image-text pairs than the WebLI-10B and CC-5.4B corpora used by the principal baselines. This comparison concerns COKE's direct training corpus, while its teacher, text initialization, and filtering models introduce additional pretrained priors. The results therefore evaluate the complete public-resource recipe rather than isolate the effect of corpus size.

Table 10: **Zero-shot image classification and retrieval.** 👁 and 👁‍🗨 mark public and nonpublic training-data sources, respectively. Entries marked with † are re-evaluated under our protocol. Other baseline results are taken from UniViTAR (Qiao et al., 2025) or the original publications. Each **Overall** column is the unweighted mean of the group it heads, namely the six classification benchmarks and the four retrieval directions. Rows form four blocks by resolution regime (fixed or native) and scale ($\leq$ 1B or $>$ 1B). Per-column **best** is bold and second-best underlined, while COKE-CLIP rows are shaded.

Method	Data Source	Res./Patch	Overall	Classification						Overall	Flickr		COCO	
				IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net		T→I	I→T	T→I	I→T
CLIP-L (Radford et al., 2021)	👁‍🗨 WIT400M	256/16	72.1	75.5	70.8	87.8	69.8	59.6	68.9	60.8	65.0	85.2	36.5	56.3
OpenCLIP-L (Ilharco et al., 2021)	👁 DC-1B	256/16	75.7	79.2	69.6	90.8	72.1	68.0	74.3	67.9	73.4	89.0	45.7	63.3
EVA02-L (Sun et al., 2023b)	👁 Merged-2B	336/14	77.5	79.8	76.2	92.7	73.0	68.1	74.9	69.9	78.0	89.6	47.9	64.2
MetaCLIP-H (Xu et al., 2023)	👁‍🗨 CC-2.5B	256/16	78.4	80.5	75.3	93.4	74.2	70.5	76.4	71.3	78.3	91.8	48.8	66.2
SigLIP-L (Zhai et al., 2023)	👁‍🗨 WebLI-10B	384/16	80.7	82.1	76.5	95.0	75.9	73.6	80.9	74.7	82.6	92.9	52.8	70.5
CLIPAv2-H (Li et al., 2023d)	👁 DC-1B	336/14	80.8	81.8	82.7	94.4	75.6	72.8	77.4	70.8	76.3	90.3	49.2	67.2
DFN-H (Fang et al., 2023)	👁‍🗨 DFN-5B	378/14	80.5	84.4	79.6	93.8	78.3	73.2	73.4	75.9	82.0	94.0	55.6	71.9
SigLIP-SO (Zhai et al., 2023)	👁‍🗨 WebLI-10B	384/16	81.7	83.1	82.5	95.8	77.2	74.5	77.0	76.0	83.0	94.3	54.2	72.4
†SigLIP2-SO (Tschannen et al., 2025)	👁‍🗨 WebLI-10B	384/16	84.6	84.2	87.0	96.1	78.5	**76.2**	85.4	77.4	84.1	94.9	56.9	73.7
†PE_{core}-L (Bolya et al., 2025)	👁‍🗨 CC-5.4B	336/14	83.9	83.5	89.1	95.2	77.9	73.4	84.5	78.8	85.5	96.5	57.1	75.9
COKE-CLIP-H	👁 COKE-1.7B	256/16	85.0	84.4	89.6	96.5	79.0	75.7	84.7	80.1	86.4	96.6	59.4	77.9
COKE-CLIP-H	👁 COKE-1.7B	384/16	85.9	84.8	92.1	**96.7**	**79.8**	76.0	86.0	**80.4**	**87.1**	96.6	59.6	**78.1**
COKE-CLIP-H	👁 COKE-1.7B	512/16	**86.1**	**84.9**	**92.7**	**96.7**	79.6	76.1	**86.4**	**80.4**	86.8	**96.7**	**60.1**	78.0
KeyeViT-SO (Team et al., 2025)	👁‍🗨 In-House	Native/16	81.6	82.7	83.3	95.2	76.8	72.6	78.7	-	-	-	-	-
UniViTAR-H (Qiao et al., 2025)	👁 Merged-1B	Native/14	82.1	82.3	86.8	94.9	76.1	71.6	81.1	76.6	84.1	95.5	55.4	71.7
Seed-ViT (Guo et al., 2025)	👁‍🗨 In-House	Native/16	82.5	83.6	85.5	95.2	77.6	74.1	79.2	-	-	-	-	-
Qwen3-ViT-SO (Bai et al., 2025)	👁‍🗨 In-House	Native/16	83.6	84.6	87.1	95.7	78.8	74.5	81.0	-	-	-	-	-
COKE-CLIP-H	👁 COKE-1.7B	Native/16	**85.4**	**84.8**	**91.6**	**96.7**	**79.5**	**75.9**	**83.8**	**80.8**	**87.4**	**96.9**	**60.8**	**78.0**
OpenCLIP-G (Ilharco et al., 2021)	👁 LAION-2B	224/14	76.2	80.1	69.3	92.1	73.6	68.9	72.8	72.8	79.6	92.9	51.4	67.4
EVA02-E (Sun et al., 2023b)	👁 Merged-2B	336/14	80.9	82.0	82.2	94.6	75.6	71.6	79.4	73.2	78.9	94.1	51.1	68.7
CLIPAv2-G (Li et al., 2023d)	👁 DC-1B	336/14	82.7	83.1	86.0	95.4	77.3	74.5	79.7	72.2	78.3	92.2	50.4	67.8
InternViT-6B (Chen et al., 2024c)	👁 InternVL-5B	224/14	82.5	83.2	83.8	95.7	77.3	74.3	80.6	75.3	81.7	94.7	54.1	70.6
EVA-8B (Sun et al., 2023a)	👁 Merged-2B	224/14	82.9	83.5	85.2	95.3	77.7	74.3	81.2	74.9	80.8	95.6	53.0	70.3
EVA-18B (Sun et al., 2023a)	👁 Merged-2B	224/14	83.6	83.8	87.3	95.7	77.9	74.7	82.2	77.5	83.3	96.7	56.2	73.6
†SigLIP2-g-1B (Tschannen et al., 2025)	👁‍🗨 WebLI-10B	384/16	86.2	85.0	90.5	96.6	80.0	**77.4**	87.5	78.1	84.2	95.8	57.4	75.2
† PE_{core}-G (Bolya et al., 2025)	👁‍🗨 CC-5.4B	448/14	86.6	**85.4**	92.6	96.5	80.2	76.5	**88.2**	78.9	85.7	96.2	58.2	75.5
COKE-CLIP-G	👁 COKE-1.7B	256/16	85.8	84.9	91.3	96.5	79.5	76.3	86.1	79.6	**86.5**	96.3	58.6	76.9
COKE-CLIP-G	👁 COKE-1.7B	384/16	86.6	85.3	93.3	96.9	80.2	77.1	86.7	**79.9**	**86.5**	96.6	59.3	**77.1**
COKE-CLIP-G	👁 COKE-1.7B	512/16	**86.8**	85.3	**93.5**	**97.0**	**80.5**	77.2	87.1	**79.9**	86.4	**97.1**	**59.6**	76.5
UniViTAR-1B (Qiao et al., 2025)	👁 Merged-1B	Native/14	83.5	82.9	89.1	95.7	77.3	73.4	82.8	76.3	83.5	95.1	55.3	71.3
COKE-CLIP-G	👁 COKE-1.7B	Native/16	**86.1**	**85.1**	**93.1**	**96.9**	**80.0**	**77.1**	**84.3**	**80.5**	**87.2**	**96.8**	**60.0**	**77.8**

C.3 Results on Zero-shot Video Classification & Retrieval

C.3.1 Evaluation Setup

We evaluate zero-shot classification on Kinetics-400 (Kay et al., 2017), UCF-101 (Soomro et al., 2012), and HMDB-51 (Kuehne et al., 2011), and retrieval on ActivityNet (Caba Heilbron et al., 2015), MSR-VTT (Xu et al., 2016), and MSVD (Chen & Dolan, 2011). Classification uses top-1 accuracy, while retrieval reports Recall@1 in both directions. The two Overall columns average the three classification benchmarks and six retrieval directions, respectively. Evaluation follows §C.2.1, using the Kinetics prompts of InternVL (Chen et al., 2024c) and the PE evaluation code's caption handling. For each clip, we uniformly sample 16 frames and group consecutive sampled frames into 8 cubes. Patch features are averaged within each cube, and cube descriptors are averaged temporally before projection into the shared embedding space (§2.1). At matched spatial patch counts, the eight cubes match an eight-frame baseline's patch-token count, although the input frame counts differ and total

computation is not necessarily equal. We evaluate fixed- and native-resolution checkpoints, with the effect of Snapshot-Replay assessed separately under matched video data and initialization in §C.6.7.

C.3.2 Results Comparison and Analysis

Video performance at the H scale. At $384/16$, COKE-CLIP-H achieves classification and retrieval averages of 74.0 and 56.0, exceeding PE_{core}-L at $336/14$ by 1.2 and 1.3 points. Classification improves on Kinetics-400 and UCF-101, while retrieval improves in both directions on ActivityNet and MSVD. HMDB-51 and MSR-VTT remain below PE-L. Both configurations use 576 spatial patch tokens per encoded unit, with COKE processing eight double-frame cubes and PE eight individual frames. The lower-resolution COKE-H configuration at $256/16$ also exceeds PE-L on both averages (73.6 and 56.6), using 2048 patch tokens per video compared with 4608. This demonstrates competitive performance with fewer patch tokens, without implying matched model capacity or total computation.

Video performance at the G scale. At $512/16$, COKE-CLIP-G achieves 75.3 classification and 55.7 retrieval, compared with 75.9 and 58.5 for PE_{core}-G. COKE-G leads on UCF-101 (92.4 vs. 90.2) and approaches PE-G on Kinetics-400 (76.0 vs. 76.4), while HMDB-51 remains lower (57.6 vs. 61.0). Within COKE-G, the 512 checkpoint gives the highest classification average, whereas the 256 checkpoint gives the highest retrieval average. Moving from 256 to 512 improves classification by 0.8 points but reduces retrieval by 1.4 points, indicating different task preferences across the progressively trained checkpoints.

Native-resolution performance. Under the reported native configurations, COKE-CLIP-H exceeds UniViTAR-0.6B by 4.9 classification and 2.0 retrieval points, while COKE-CLIP-G exceeds UniViTAR-1B by 5.7 and 2.8 points. Both COKE models improve all three classification benchmarks. On MSVD, H improves V2T and T2V retrieval by 8.2 and 5.3 points, while G improves them by 9.3 and 5.0 points. These results extend the observed gains to native-resolution video evaluation, with particularly strong improvements on MSVD.

C.4 Results on Image Classification by Frozen-Encoder Probing

C.4.1 Evaluation Setup

Frozen-encoder probing isolates the quality of the visual representation from that of the image-text alignment, since the backbone receives no gradient and only a classification head is trained on the ImageNet-1K (Deng et al., 2009) training set. Following the convention of the probing literature (Chen et al., 2024c; Qiao et al., 2025), we evaluate on the ImageNet-1K validation set together with ImageNet-ReaL (Beyer et al., 2020), ImageNet-V2 (Recht et al., 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019). Every COKE entry uses an *attentive* head, namely a cross-attention layer with randomly initialized queries introduced by AIMv2 (Fini et al., 2024) and adopted by both UniViTAR (Qiao et al., 2025) and PE_{core} (Bolya et al., 2025). The **Classifier** column of Table 12 records the head used by each baseline, and we draw every comparison in §C.4.2 between attentive entries and quote the linear ones for reference only. We fit a separate head for every entry with SGD at a learning rate of 0.05, a batch size of 128, and 10 epochs under the same image transform as COKE training, probing COKE at fixed resolutions of 256, 384, and 512 as well as at native resolution.

C.4.2 Results Comparison and Analysis

Attentive probing at matched patch-token budgets. Table 12 evaluates frozen visual features. At 576 patch tokens, COKE-ViT-H reaches 87.3 overall, compared with 86.3 for SigLIP 2-SO. At this budget, COKE-ViT-G reaches 87.9, compared with 87.6 for SigLIP 2-g-1B. Both comparisons use attentive heads. The largest gains occur on ImageNet-A, where H and G exceed their SigLIP 2 counterparts by 5.3 and 2.9 points. On ImageNet-Sketch, however, they remain lower by 1.6 and 2.1 points. The higher overall averages therefore reflect uneven improvements across evaluation sets, with particularly strong results on ImageNet-A.

Resolution effects. The COKE probing average increases with fixed input resolution, reaching 87.5 for H and 88.0 for G at 1024 patch tokens. Native-resolution averages are 87.1 and 87.7. These results complement the zero-shot evaluations.

Table 11: **Zero-shot video classification and retrieval.** Most baselines are quoted from the original publications, whereas † marks our own re-evaluation. Each **Overall** column is the unweighted mean of the group it heads, namely the three classification benchmarks in blue and the six retrieval directions in green. Rows form four blocks by resolution regime (fixed or native) and scale ($\leq$1B or $>$ 1B). The last three rows are set in gray because they are dedicated video models. **Frames** 16 (8) denotes 16 uniformly sampled frames grouped into 8 temporal cubes, each containing p_t=2 consecutive frames in the sampled sequence. Per-column **best** is bold and second-best underlined.

Method	Res./Patch	Frames	Overall	Classification			Overall	ANet		MSR-VTT		MSVD	
				K400	UCF	HMDB		V→T	T→V	V→T	T→V	V→T	T→V
OpenCLIP-H (Ilharco et al., 2021)	224/14	16	62.0	61.7	72.5	51.6	43.5	36.1	38.9	34.5	38.9	63.3	49.4
DFN-H (Fang et al., 2023)	378/14	16	62.9	63.8	76.7	48.2	46.2	39.7	42.9	36.1	39.6	66.6	52.4
SigLIP-SO (Zhai et al., 2023)	384/16	16	67.3	66.8	83.0	52.1	47.5	36.6	39.3	37.5	41.1	75.5	54.7
†SigLIP2-SO (Tschannen et al., 2025)	384/16	8	68.4	67.0	87.4	50.9	46.9	34.9	35.8	36.0	42.5	76.9	55.3
† PE_{core}-L (Bolya et al., 2025)	336/14	8	72.8	72.9	87.1	**58.5**	54.7	42.8	45.4	**50.1**	**50.3**	82.5	57.1
COKE-CLIP-H	256/16	16 (8)	73.6	74.1	89.4	57.3	**56.6**	**47.4**	**49.1**	49.1	50.2	**84.2**	59.4
COKE-CLIP-H	384/16	16 (8)	**74.0**	**74.7**	89.8	57.6	56.0	47.1	47.8	48.4	49.8	83.4	**59.7**
COKE-CLIP-H	512/16	16 (8)	**74.0**	74.5	**89.9**	57.7	55.5	45.9	48.2	46.6	49.6	83.4	59.4
UniViTAR-0.6B (Qiao et al., 2025)	Native/14	2∼32	68.6	67.6	82.9	55.2	54.9	**48.7**	**51.5**	48.6	**50.2**	75.8	54.3
COKE-CLIP-H	Native/16	16 (8)	**73.5**	**74.0**	**88.9**	**57.5**	**56.9**	48.1	49.8	**50.1**	49.9	**84.0**	**59.6**
OpenCLIP-G (Ilharco et al., 2021)	224/14	16	64.2	63.2	76.2	53.4	46.0	36.7	41.4	36.9	41.8	67.5	51.5
EVA01-g (Sun et al., 2023a)	224/14	16	62.8	63.4	72.1	52.9	45.5	37.0	40.1	37.2	40.1	67.6	50.8
InternViT-6B (Chen et al., 2024c)	224/14	8	-	69.1	-	-	-	-	-	42.4	46.3	-	-
† PE_{core}-G (Bolya et al., 2025)	448/14	8	**75.9**	**76.4**	90.2	**61.0**	**58.5**	**51.6**	**53.3**	**50.2**	**51.2**	**85.2**	59.7
COKE-CLIP-G	256/16	16 (8)	74.5	75.2	91.6	56.6	57.1	47.7	49.7	49.9	51.0	84.5	59.8
COKE-CLIP-G	384/16	16 (8)	75.0	75.9	91.9	57.3	56.7	47.9	49.9	48.3	50.1	83.4	**60.3**
COKE-CLIP-G	512/16	16 (8)	75.3	76.0	**92.4**	57.6	55.7	46.7	48.4	47.7	48.5	82.8	59.8
UniViTAR-1B (Qiao et al., 2025)	Native/14	2∼32	69.0	68.6	81.0	57.3	54.0	47.8	**49.6**	48.3	47.6	75.5	55.2
COKE-CLIP-G	Native/16	16 (8)	**74.7**	**75.2**	**91.1**	**57.8**	**56.8**	**48.1**	48.7	**48.8**	**50.1**	**84.8**	**60.2**
VideoCoCa-g (Yan et al., 2022)	224	16	72.4	72.0	86.6	58.7	-	33.0	34.5	64.7	34.4	-	-
VideoPrism-g (Zhao et al., 2024)	288	16	-	76.4	-	-	-	50.3	52.7	71.0	39.7	-	-
InternVideo2-6B (Wang et al., 2024b)	224	8	-	-	-	-	62.0	56.5	63.2	53.7	55.9	83.1	59.3

C.5 Results on Multimodal Understanding

We evaluate the multimodal transfer performance of COKE-ViT under a shared LLaVA-NeXT fine-tuning recipe, with the vision encoder updated during instruction tuning. The baselines include SigLIP (Zhai et al., 2023), MetaCLIP 2 (Chuang et al., 2025), PE_{core} (Bolya et al., 2025), and SigLIP 2 (Tschannen et al., 2025). Table 3 compares encoder configurations under Square inputs and studies Split-by-Tile and Native configurations at larger visual token budgets.

C.5.1 Evaluation Setup

Unified Training Recipe. We compare vision towers and input paradigms under a shared multimodal training recipe. All models adopt the same two-stage LLaVA-NeXT recipe with a two-layer MLP projector and Qwen2.5-7B-Instruct. Stage-1 aligns the projector alone on the LLaVA-v1.5 pretraining data and Stage-2 runs instruction tuning on the LLaVA-NeXT raw-format mixture with all parameters unfrozen. We use a global batch size of 128, a cosine schedule with warmup ratio 0.03, and learning rates of 1×10^{-3} in Stage-1 and 1×10^{-5} for all three modules in Stage-2. Training is in BF16 with gradient checkpointing, under a context length of 8192 in Stage-1 and 32768 in Stage-2. Experiments reusing a completed alignment stage load the Stage-1 projector of the matching fixed-resolution tower, so the *Square* and *Split-by-Tile* runs for a given fixed-resolution tower share their Stage-1 initialization.

Input paradigms and token accounting. *Square* resizes each image to a fixed square and encodes it once. The Square blocks match patch-token budgets of 576 or 1024. COKE additionally passes its four register and eight reference tokens through the projector to the language model, giving 588 or 1036 visual tokens in total. *Split-by-Tile* selects a canvas from aspect ratios $(1,2)$, $(2,1)$, $(2,2)$, $(3,1)$, and $(1,3)$, partitions it into fixed-size tiles, and retains a square global view. Tile features are

Table 12: **Frozen-encoder probing on ImageNet-1K and its variants**. Most baseline numbers are quoted as reported by UniViTAR (Qiao et al., 2025), whereas entries marked with † are reproduced by us. Rows form four blocks by resolution regime (fixed or native) and scale ($\leq$ 1B or $>$ 1B), and comparisons should account for the patch-token count implied by Res./Patch. Parentheses give patch-token counts for fixed-resolution inputs.

Method	Classifier	Res./Patch	Overall	ImageNet Variants					
				IN-1K	IN-Real	IN-V2	IN-A	IN-R	IN-S
CLIP-H (Radford et al., 2021)	Linear	224/14 (256)	-	84.4	88.4	75.5	-	-	-
AIMv2-H (Fini et al., 2024)	Attentive	224/14 (256)	-	87.5	-	-	-	-	-
DFN-H (Fang et al., 2023)	Linear	378/14 (729)	81.6	87.3	90.4	78.8	74.8	90.3	68.3
SigLIP-SO (Zhai et al., 2023)	Attentive	384/16 (576)	-	87.3	-	-	-	-	-
†SigLIP2-SO (Tschannen et al., 2025)	Attentive	384/16 (576)	86.3	88.9	91.0	82.3	87.3	94.7	**73.9**
COKE-ViT-H	Attentive	256/16 (256)	86.8	89.2	91.2	82.4	90.3	95.2	72.4
COKE-ViT-H	Attentive	384/16 (576)	87.3	89.6	**91.3**	**82.9**	92.6	**95.4**	72.3
COKE-ViT-H	Attentive	512/16 (1024)	**87.5**	**89.7**	**91.3**	**82.9**	**93.2**	**95.4**	72.4
NaViT-L (Dehghani et al., 2023b)	Linear	Native/14	-	76.0	-	-	65.5	-	-
UniViTAR-0.6B (Qiao et al., 2025)	Attentive	Native/14	84.8	88.3	90.7	81.0	87.3	92.5	68.8
COKE-ViT-H	Attentive	Native/16	**87.1**	**89.8**	**91.0**	**82.7**	**92.0**	**94.9**	**72.3**
DINOv2-g (Oquab et al., 2023)	Linear	224/14 (256)	78.6	86.5	89.6	78.4	75.9	78.8	62.5
InternViT-6B (Chen et al., 2024c)	Linear	224/14 (256)	82.5	88.2	90.4	79.9	77.5	89.8	69.1
†DINOv3-7B (Siméoni et al., 2025)	Attentive	224/16 (196)	86.0	89.3	91.1	82.7	89.6	91.6	71.6
†SigLIP2-g-1B (Tschannen et al., 2025)	Attentive	384/16 (576)	87.6	89.6	91.0	83.4	90.4	95.5	**75.8**
PE_{core}-G (Bolya et al., 2025)	Attentive	448/14 (1024)	-	89.8	-	-	-	-	-
COKE-ViT-G	Attentive	256/16 (256)	87.7	89.8	91.2	83.4	91.9	95.7	74.1
COKE-ViT-G	Attentive	384/16 (576)	87.9	90.1	**91.3**	83.5	93.3	95.7	73.7
COKE-ViT-G	Attentive	512/16 (1024)	**88.0**	**90.3**	**91.3**	**83.6**	**93.6**	**95.8**	73.7
UniViTAR-1B (Qiao et al., 2025)	Attentive	Native/14	86.0	89.2	**91.0**	81.7	90.1	93.6	70.6
COKE-ViT-G	Attentive	Native/16	**87.7**	**90.2**	90.9	**83.3**	**93.0**	**95.1**	**73.9**

reassembled on a two-dimensional grid with a learnable newline token after each patch row. COKE includes the register and reference tokens of the global view and every tile in the language-model input. *Native* uses the native-resolution checkpoint and encodes the image once at its aspect ratio. For each image, we first calculate the complete visual token count of the corresponding COKE Split-by-Tile configuration, including all register, reference, and newline tokens. We then retain one set of register and reference tokens and allocate the remaining budget to native patch tokens. Native and its corresponding COKE tiled configuration therefore match the total visual tokens supplied to the language model for each image. They use different encoder checkpoints and attention layouts, so this is a comparison of configurations rather than a single-variable input ablation. Matched language-model token counts also do not imply equal vision-encoder FLOPs.

Protocol and benchmarks. All numbers come from a single lmms-eval (Zhang et al., 2025) pipeline with the Qwen conversation template. We evaluate 19 tasks in three groups, namely general multimodal understanding (RealWorldQA (xAI, 2024), SEEDBench-Image (Li et al., 2023a), MME (Fu et al., 2023), MMBench (Liu et al., 2023), VizWiz (Gurari et al., 2018), RefCOCO (Yu et al., 2016), POPE (Li et al., 2023f), HallusionBench-Image (Guan et al., 2024)), chart, document, and OCR understanding (AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), OCRBench (Liu et al., 2024b), OCRBench-V2 (Fu et al., 2024)), and reasoning (GQA (Hudson & Manning, 2019), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024a), MathVista-Testmini (Lu et al., 2023)). All scores are normalized to a 0 to 100 scale, HallusionBench-Image reports the mean of aAcc, qAcc, and fAcc, and MME is the sum of its perception and cognition scores divided by 28.

C.5.2 Results Comparison and Analysis

Transfer under Square inputs. At 576 patch tokens, COKE-ViT-H achieves 55.19 overall, compared with 54.66 for SigLIP 2-SO. General understanding and reasoning improve by 1.68 and 0.35 points, while OCR/Chart/Document performance decreases by 0.69 points. HallusionBench (28.7 vs. 18.9) contributes approximately 0.52 points to H's 0.53-point overall gain. At 1024 patch tokens, COKE-ViT-G exceeds SigLIP 2-g by 0.95 overall, with gains of 0.85, 1.48, and 0.19 points across the three groups. Its OCR/Chart/Document improvement extends across all seven benchmarks. These comparisons match patch-token counts, while COKE supplies 12 additional non-patch tokens to the language model.

Tiled high-resolution inputs. Moving from Square to Split-by-Tile increases the overall score by 4.08 points for H and 1.21 points for G. The gains are concentrated in OCR/Chart/Document tasks, whose group averages improve by 9.07 and 4.05 points, respectively. DocVQA increases by 20.1 points for H and 10.7 points for G. Improvements are not uniform across task groups, as G's general-understanding average decreases by 0.73 points. These results show that the larger visual token budget primarily benefits text- and document-oriented tasks under the tested configurations.

Native and tiled configurations. At matched total visual token counts per image, Native shows small average improvements over the corresponding COKE tiled configurations, with gains of 0.18 for H (59.45 vs. 59.27) and 0.46 for G (60.31 vs. 59.85). The group-level patterns differ across scales. For H, a 1.22-point improvement in OCR/Chart/Document performance is partly offset by lower general-understanding and reasoning averages. For G, all three group averages improve, with HallusionBench contributing approximately 0.34 points to the 0.46-point overall gain. Native improves DocVQA at both scales, while tiling remains stronger on RealWorldQA and, for G, on both OCRBench variants. In terms of input processing, Native provides a unified path by encoding one continuous image without separate tile encoding and feature reassembly.

Combined scale and resolution changes. COKE-G at 512 Square exceeds COKE-H at 384 Square, with the largest group-level difference in OCR/Chart/Document tasks. Since both model size and resolution change, this comparison does not isolate their individual effects.

C.6 Ablation Study

We study learnable tokens, distillation pathways, teacher choice, text initialization, video fine-tuning, and vocabulary-based data selection. The image-side experiments in Tables 13 to 18 use COKE-ViT-H at R=256, with 1.6B image-text samples per run (one epoch, §C.1.1). Comparisons match training budgets, with configuration differences specified below. Only the separate VRP convergence experiment uses COKE-ViT-SO400M-384. We select register/reference token counts and Snapshot-Replay hyperparameters using aggregate performance on the reported ablation benchmarks, without a separate validation set. These configurations are selected among tested settings, not independently validated global optima. The video strategy and hyperparameter experiments use different sample budgets, specified below. The vocabulary experiments in §C.6.9 use a separate training configuration with 1B image-text samples per run.

C.6.1 Register and Reference Branches

Table 13 evaluates register tokens and reference distillation independently and jointly. Registers participate in internal token interactions, while reference tokens provide dedicated outputs for teacher supervision. Registers increase the overall score from 74.73 to 75.42, while reference distillation alone reaches 78.54. Their combination obtains 79.43 and leads on nine of the ten metrics. Adding registers to reference distillation further improves the overall score by 0.89, supporting their utility even with teacher supervision. Both components contribute under this configuration, although the small difference between additive and joint gains does not establish a particular interaction mechanism.

C.6.2 Number of Reference Tokens

Table 14 varies the number of reference tokens while keeping four register tokens. The zero-token setting disables reference distillation, whereas nonzero settings assess token-count sensitivity with supervision applied to pooled reference features. Pooling produces a single distillation embedding,

Table 13: **Effect of learnable tokens in the visual backbone.** We add both token types separately and jointly to a plain contrastive encoder, where ***[REG]*** denotes register tokens, ***[REF]*** denotes reference tokens with distillation, and **Overall** is the unweighted mean of six image classification accuracies and four retrieval Recall@1 scores from Table 10.

[REG]	*[REF]*	Overall	Classification						Flickr		COCO	
			IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
✖	✖	74.73	79.2	66.9	91.6	72.6	68.8	74.1	79.1	92.1	52.3	70.6
✔	✖	75.42	79.5	70.7	92.1	72.9	68.8	74.9	79.7	92.7	52.3	70.6
✖	✔	78.54	82.1	80.1	94.9	76.2	73.0	79.7	**81.0**	93.7	53.3	71.4
✔	✔	**79.43**	**82.6**	**83.6**	**95.3**	**77.0**	**73.7**	**81.9**	80.6	**93.8**	**53.6**	**72.2**

so increasing token count does not introduce additional teacher targets. A single reference token increases the overall score from 75.42 to 79.32. All nonzero settings lie between 79.28 and 79.43, indicating limited sensitivity within this range. Eight and sixteen tokens score 79.43 and 79.40, respectively. We select eight because it uses fewer tokens and has a slightly higher observed average, without interpreting the 0.03 difference as a stable performance advantage. Increasing the count from eight to sixteen raises the classification mean from 82.35 to approximately 82.57 but lowers retrieval from 75.05 to 74.65, showing that more reference tokens do not consistently improve both task groups. Most of the overall improvement is observed when enabling reference distillation, with limited additional benefit from increasing the number of reference tokens.

Table 14: **Effect of the number of reference tokens.** Register tokens are fixed at four while the number of reference queries that receive the distillation signal varies from zero to sixteen. **Overall** follows the same definition as in Table 13.

#[REF]	Overall	Classification						Flickr		COCO	
		IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
0	75.42	79.5	70.7	92.1	72.9	68.8	74.9	79.7	92.7	52.3	70.6
1	79.32	82.5	83.9	95.3	76.8	73.6	**82.1**	80.4	93.6	53.5	71.5
2	79.28	82.4	83.3	95.2	76.4	73.4	82.0	80.8	**93.9**	53.3	72.1
4	79.35	82.6	83.7	95.3	76.8	73.2	81.7	**81.0**	93.8	**53.7**	71.7
8	**79.43**	82.6	83.6	95.3	**77.0**	**73.7**	81.9	80.6	93.8	53.6	**72.2**
16	79.40	**82.9**	**84.5**	**95.5**	**77.0**	**73.7**	81.8	80.2	93.3	53.6	71.5

C.6.3 Number of Register Tokens

Table 15 varies register count with reference distillation disabled, isolating their contribution to contrastive learning without auxiliary visual supervision. Registers are excluded from patch pooling, so their contribution arises through token interactions within the shared backbone. Four registers yield the highest observed overall score, 75.42, compared with 74.73 without registers. Much of the gain occurs on ImageNet-A (66.9 to 70.7). Eight registers score 75.28, so we retain four as the selected setting in this sweep. The small differences among nonzero settings do not establish a general optimum.

Table 15: **Effect of the number of register tokens.** Reference distillation is switched off to isolate the register mechanism, while the register count varies from zero to eight. **Overall** follows the same definition as in Table 13.

#[REG]	Overall	Classification						Flickr		COCO	
		IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
0	74.73	79.2	66.9	91.6	72.6	68.8	74.1	79.1	92.1	52.3	70.6
1	75.17	**79.5**	70.0	92.0	**73.2**	**68.9**	74.2	79.2	92.0	52.3	70.4
2	75.36	**79.5**	70.2	**92.1**	73.1	68.8	74.5	79.2	**93.3**	52.2	**70.7**
4	**75.42**	**79.5**	**70.7**	**92.1**	72.9	68.8	74.9	**79.7**	92.7	52.3	70.6
8	75.28	**79.5**	70.6	**92.1**	72.8	**68.9**	**75.1**	79.0	92.0	**52.5**	70.3

C.6.4 Reference Injection Pathways

Table 16 compares pooled-patch and reference-token distillation under the same teacher, distillation weight, and sample budget. All three configurations use four register tokens. The no-distillation and pooled-patch configurations use no reference tokens, while the reference-token configuration adds eight. Both variants use a LayerNorm-linear projector followed by L2 normalization and the same temperature-scaled KL objective, applied to pooled patch or reference-token features. Image-text alignment uses a separate projection of pooled patch features in both variants. Relative to no distillation, pooled-patch supervision improves all six classification scores (mean 76.48 to 79.55) but lowers all four retrieval scores (mean 73.83 to 69.85). Reference-token supervision outperforms both alternatives on all ten metrics, reaching classification and retrieval means of 82.35 and 75.05, respectively, with gains of 2.80 and 5.20 points over pooled-patch supervision. These results support the tested reference-token design, while the contributions of added tokens and the distillation output choice are not isolated. At 400M samples, reference-token distillation outperforms all tested pooled-patch weights (Appendix D.5).

Table 16: **Distillation configurations.** All rows use four register tokens. *None* and *Patch tokens* use no reference tokens, while *Reference tokens* uses eight. Both variants apply the same projector architecture and loss to pooled patch or reference-token features, with a separate head for image-text alignment. **Overall** follows the same definition as in Table 13.

Injection Pathway	Overall	Classification						Flickr		COCO	
		IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
None	75.42	79.5	70.7	92.1	72.9	68.8	74.9	79.7	92.7	52.3	70.6
Patch tokens (pooled)	75.67	80.5	78.3	94.2	74.2	71.3	78.8	75.5	90.2	48.4	65.3
Reference tokens	**79.43**	**82.6**	**83.6**	**95.3**	**77.0**	**73.7**	**81.9**	**80.6**	**93.8**	**53.6**	**72.2**

C.6.5 Reference Teacher

Table 17 changes the frozen teacher backbone from DINOv2-g to DINOv3-7B while keeping the student backbone configuration and sample budget fixed. The student distillation projector's output dimension is adjusted to match each teacher's feature dimension. The overall score increases from 77.65 to 79.43, with improvements across all ten metrics. The classification mean increases from 80.50 to 82.35, while retrieval improves from approximately 73.38 to 75.05. ObjectNet shows the largest individual gain of 3.6 points. Teacher supervision through reference outputs influences the shared contrastive representation without changing the student backbone architecture. The gains across both groups show that the tested teacher replacement benefits both classification and cross-modal retrieval.

Table 17: **Effect of the reference teacher.** The student backbone, reference-token count, register-token count, and training schedule are fixed. The student distillation projector's output dimension matches the selected teacher. **Overall** follows the same definition as in Table 13.

Teacher	Overall	Classification						Flickr		COCO	
		IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
DINOv2-g	77.65	82.0	81.2	93.6	75.8	72.1	78.3	79.2	92.9	52.2	69.2
DINOv3-7B	**79.43**	**82.6**	**83.6**	**95.3**	**77.0**	**73.7**	**81.9**	**80.6**	**93.8**	**53.6**	**72.2**

C.6.6 Text Branch Initialization and Update Policy

Table 18 compares text-tower initialization and update policies. The trainable configurations isolate the benefit of initialization. Under the same 1.6B-sample budget, a pretrained and trainable text tower improves the overall score from 72.52 to 74.44 relative to random initialization. Freezing the pretrained tower yields 74.73 overall, compared with 74.44 when it is updated. This small aggregate difference conceals a task trade-off. Freezing improves the classification mean by approximately 1.28 points but reduces retrieval by 1.20 points. This trade-off is consistent across benchmarks. Freezing performs better on all six classification metrics, whereas updating performs better on all four retrieval metrics. Freezing preserves the pretrained text representation as a fixed alignment target while the

visual encoder adapts. We adopt freezing during VRP for its observed aggregate performance and reduced training computation. These results support the choice under the tested budget without establishing a general advantage over text-tower fine-tuning.

Table 18: **Effect of text-branch initialization and update policy.** All rows share the common ablation protocol and differ only in whether the text encoder starts from pretrained weights and whether it is updated in the 1.6B-sample run.

Text Encoder		Overall	Classification						Flickr		COCO	
Pretrained	Tuned		IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net	T→I	I→T	T→I	I→T
✖	✔	72.52	77.6	64.1	90.3	70.7	67.0	71.1	76.5	91.7	48.9	67.3
✔	✖	**74.73**	**79.2**	**66.9**	**91.6**	**72.6**	**68.8**	**74.1**	79.1	92.1	52.3	70.6
✔	✔	74.44	78.5	64.9	90.6	72.3	68.1	71.1	**79.9**	**94.0**	**52.9**	**72.1**

C.6.7 Video Fine-tuning Strategies

We compare continuous video-only fine-tuning, EMA (Tarvainen & Valpola, 2017), Lookahead (Zhang et al., 2019), WiSE-FT (Wortsman et al., 2022b), and Snapshot-Replay using the same image-pretrained COKE-ViT-H-256 checkpoint. Each fine-tuning run uses the 64M video-text collection for 10 epochs, totaling 640M samples, without replaying image-text pairs. EMA uses a decay coefficient of 0.99. Lookahead uses a synchronization interval of 5 optimizer steps and an interpolation coefficient of 0.5. WiSE-FT reuses the continuous video-only training run and interpolates its endpoint with the initial checkpoint using $\beta = 0.5$ in the main comparison. An additional interpolation sweep is reported in Appendix D.4. Snapshot-Replay is evaluated at its final fused weights ϕ_N. Table 5 shows continuous video-only fine-tuning raising the video average from 56.23 to 61.20 while reducing the image average from 81.55 to 81.21. EMA and Lookahead test whether averaging or interpolation along the training trajectory improves retention. Changes from continuous fine-tuning are small, with image averages below initialization. WiSE-FT raises the image average to 82.79, although its video average of 61.01 is slightly below continuous fine-tuning. Snapshot-Replay resumes video optimization after each interpolation with the fixed image checkpoint, repeatedly incorporating pretrained weights. Its image and video averages of 83.02 and 62.23 are the highest observed, exceeding continuous fine-tuning by 1.81 and 1.03 points, respectively. Against WiSE-FT at $\beta = 0.5$, the image margin is small (0.23 points), while video performance improves by 1.22 points. It also exceeds the best image and video averages across the WiSE-FT sweep by 0.23 and 0.46 points, respectively (Appendix D.4). These results support periodic interpolation to improve image retention and video performance under the tested initialization and training budget.

C.6.8 Snapshot-Replay Hyperparameters

We evaluate the burst length K and fusion weight α using the same image-pretrained COKE-ViT-H-256 checkpoint as §C.6.7. At a fixed training budget, K controls both the length of uninterrupted video optimization and the frequency of fusion, while α controls the contribution of the fixed image-pretrained checkpoint. The two parameters therefore jointly govern the adaptation trajectory. To reduce training cost, each run uses 2 epochs over the 64M video-text collection, totaling 128M samples. These scores should therefore be compared within Table 19, rather than directly with the ten-epoch strategy results. At $\alpha = 0.75$, increasing K from 1 to 10 improves the image average from 82.55 to 82.90 but reduces the video average from 61.76 to 61.24. Extending the burst to $K = 100$ lowers both averages relative to $K = 10$. At $K = 10$, $\alpha = 0.75$ yields the highest image score, whereas $\alpha = 0.01$ yields the highest video score. Increasing the snapshot weight from $\alpha = 0.75$ to 0.99 lowers both averages, showing that stronger anchoring does not necessarily improve image retention and reduces video performance in this sweep. We select $K = 10$ and $\alpha = 0.75$ for their image performance and competitive video score within these sweeps.

C.6.9 Effectiveness of Vocabulary-Based Data Selection

Experimental setup. We evaluate vocabulary-based data selection under a separate training configuration. All six runs train COKE-ViT-H at 256 resolution on 1B image-text samples, with a randomly initialized visual backbone and a text tower initialized from SigLIP 2-SO400M-384, frozen

Table 19: **Snapshot-Replay hyperparameters.** All runs use COKE-ViT-H-256, the same image-pretrained initialization, and 128M video-text samples (2 epochs). One parameter varies while the other remains at its selected value, $K = 10$ or $\alpha = 0.75$. Image and video averages follow Table 5.

K	**Overall**$_{\text{Image}}$	**Overall**$_{\text{Video}}$
1	82.55	**61.76**
10	**82.90**	61.24
100	82.51	59.71

(a) Burst length, with $\alpha = 0.75$.

α	**Overall**$_{\text{Image}}$	**Overall**$_{\text{Video}}$
0.99	81.81	57.59
0.75	**82.90**	61.24
0.01	82.16	**61.61**

(b) Snapshot weight, with $K = 10$.

throughout. Without VRP, training uses only the image-text contrastive objective. Optimizer settings, learning-rate schedules, batch size, data augmentation, and evaluation protocols are identical. The sources are released DataComp-1B and the approximately 4.5B-pair public pool obtained by quality filtering and deduplicating the original 18.7B collection (§2.4). On DataComp-1B, we compare full-data training with random and MetaCLIP selection, using $\texttt{max_match} = 10$K for MetaCLIP. On the public pool, we compare MetaCLIP, COKE-K, and COKE-V at $\texttt{max_match} = 40$K. All vocabulary-based selections share one concept-matching and sampling procedure. These selection settings are specific to this ablation, with main-corpus settings given in §2.4. Table 6 summarizes selected pool sizes and training epochs, with the four-epoch sampling rule specified in its caption. Following Table 10, Overall$_{\text{CLS}}$ and Overall$_{\text{RET}}$ average the six classification and four retrieval metrics, respectively, while Overall$_{\text{Image}}$ averages all ten.

Selection on DataComp-1B. At the same cumulative budget of 1B samples, MetaCLIP selection outperforms random sampling by 1.12, 1.20, and 1.15 points in classification, retrieval, and overall averages, respectively. Full-data training achieves 70.68 overall, trailing MetaCLIP by 0.50 points and exceeding random selection by 0.65. These results support vocabulary-based selection under the tested budget without establishing a general advantage from using fewer unique training pairs.

Vocabulary comparison on the public pool. Replacing MetaCLIP's vocabulary with COKE-K improves classification, retrieval, and overall averages by 0.73, 0.89, and 0.80 points, respectively. With the candidate pool, sampling cap, selected pool size, and training schedule fixed, this supports COKE-K vocabulary selection in this configuration. COKE-V achieves the highest classification and overall averages in Table 6, reaching 74.03 and 73.26. Relative to MetaCLIP on the same public pool, its classification, retrieval, and overall gains are 3.72, 0.62, and 2.48 points, respectively. COKE-K retains the highest retrieval average of 72.38, compared with 72.11 for COKE-V. These results support both vocabulary-based selection strategies, with COKE-V's advantage over MetaCLIP concentrated in classification. COKE-V uses a smaller selected pool with repeated sampling, so its gains do not isolate vocabulary choice from data reuse.

D Additional Analyses and Results

D.1 Additional Model Comparisons

Figure 5 supplements the results in Table 10 by comparing classification performance against direct training corpus size. The classification score is the unweighted mean accuracy across the six benchmarks described in §C.2.1. Corpus size counts unique image-text pairs rather than cumulative training samples. The COKE points use the 512-resolution checkpoints, with the 1.68B-pair corpus rounded to 1.7B in the plot.

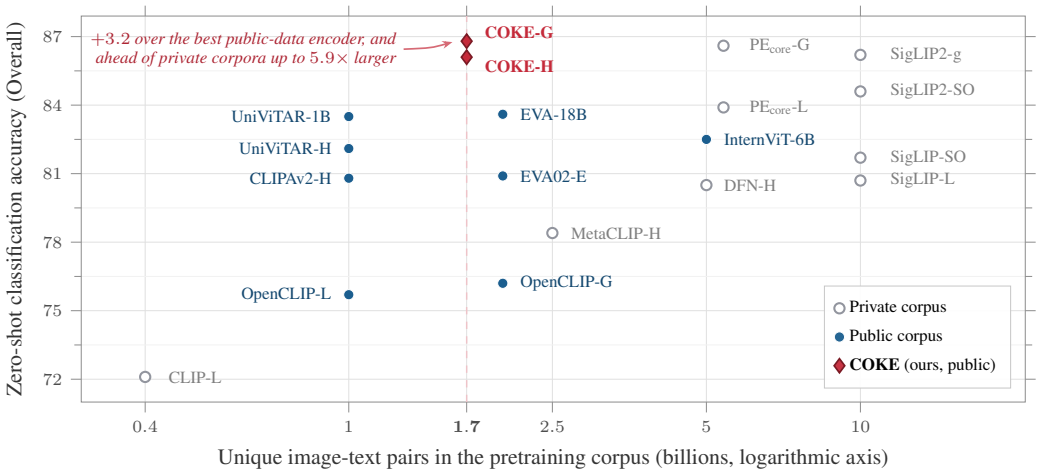


Figure 5: **Zero-shot classification against corpus size.** Each point is one encoder from Table 10, positioned by the number of unique pairs its corpus contains and by its **Overall** classification score. Filled and hollow markers indicate public and private corpora, respectively. COKE-CLIP-H and COKE-CLIP-G are trained on the COKE-1.7B corpus assembled in §2.4, and at its best configuration COKE-CLIP-G reaches 86.8, exceeding the listed baselines. Corpus sizes describe direct image-text training data and exclude the pretraining data of auxiliary models. The figure does not isolate corpus size from other differences in the training recipes.

D.2 Early Sample Efficiency of Visual Reference Pretrain

To measure early sample efficiency, we compare two COKE-ViT-SO400M-384 runs, each trained on 1.6B image-text samples from COKE-1.7B. Both use randomly initialized visual backbones, frozen text towers initialized from SigLIP 2-SO400M-384, and identical optimization settings. Both retain four register and eight reference tokens. The VRP run applies teacher supervision throughout training, while the baseline disables the teacher branch and distillation loss. We report mean zero-shot classification accuracy over the six benchmarks in Table 10. This model configuration is exclusive to the convergence experiment.

In Fig. 6, VRP reaches 76.4 after 0.6B samples, exceeding the baseline's 75.5 after 1.6B samples. This corresponds to approximately 2.7× fewer samples for comparable accuracy. At 1.6B samples, the VRP run obtains 82.6, compared with 75.5 without VRP. SigLIP 2-SO400M (84.6 after 40B samples) provides an external reference. Its different data and training recipe preclude a controlled efficiency comparison. The VRP sample-efficiency ratio excludes additional teacher forward computation and does not imply a wall-clock or FLOP speedup.

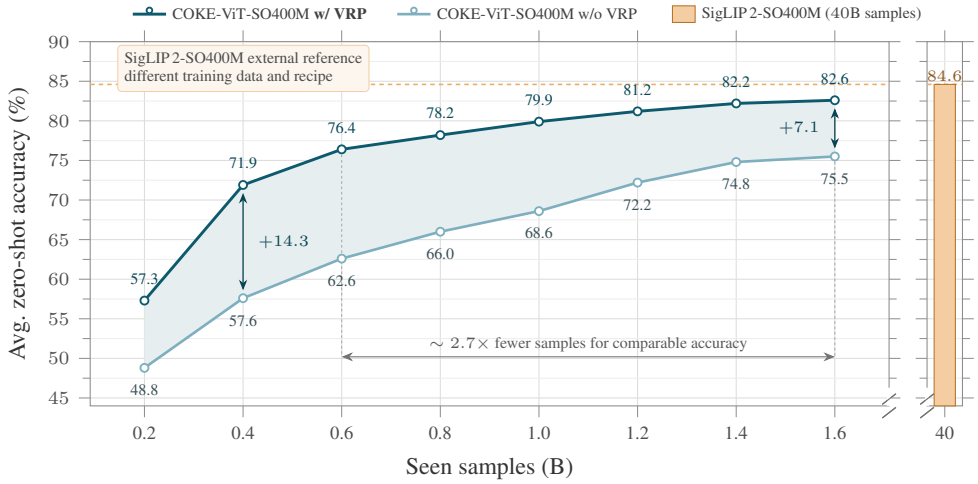


Figure 6: **Early sample efficiency of Visual Reference Pretrain.** Both controlled runs use COKE-ViT-SO400M-384 and report mean zero-shot classification accuracy over the six benchmarks in Table 10. VRP reaches 76.4 after 0.6B samples, compared with 75.5 after 1.6B without VRP. The approximately 2.7× ratio concerns sample count and excludes teacher computation. SigLIP 2-SO400M is an external reference with different data and training settings.

D.3 Interpretation of the Fusion Step

For $0 \leq \alpha \leq 1$, each Snapshot-Replay fusion has the equivalent form

$$\phi_n = \arg\min_{\theta} \left\{ \frac{1-\alpha}{2} \|\theta - \psi_n\|_2^2 + \frac{\alpha}{2} \|\theta - \theta_0\|_2^2 \right\}, \tag{4}$$

where optimization is over the trainable parameters. Setting the gradient to zero gives $\phi_n = \alpha\theta_0 + (1-\alpha)\psi_n$. Thus, fusion balances proximity to the video-trained endpoint and the fixed image-pretrained checkpoint. This interpretation applies to the fusion step alone and does not make the full AdamW training trajectory equivalent to optimizing a fixed regularized objective.

Parameter displacement. The fixed interpolation target bounds parameter displacement under a bounded-update assumption. Writing $d_n = \phi_n - \theta_0$ and $\delta_n = \psi_n - \phi_{n-1}$, the update becomes $d_n = (1-\alpha)(d_{n-1} + \delta_n)$. For $0 < \alpha \leq 1$ and $\|\delta_j\| \leq D$ at every round, this gives

$$\phi_n - \theta_0 = \sum_{j=1}^{n} (1-\alpha)^{n-j+1} \delta_j \,, \qquad \|\phi_n - \theta_0\| \leq \frac{(1-\alpha)[1-(1-\alpha)^n]}{\alpha} D \leq \frac{1-\alpha}{\alpha} D. \tag{5}$$

The bound limits displacement from the image-text checkpoint under this assumption but does not guarantee image-performance retention. We therefore evaluate image retention and video adaptation empirically.

D.4 WiSE-FT Interpolation Sweep

We extend the WiSE-FT baseline in Table 5 by interpolating the same image-pretrained checkpoint θ_0 and continuous video fine-tuning endpoint ψ:

$$\theta_\beta = (1-\beta)\theta_0 + \beta\psi, \qquad \beta \in \{0, 0.1, \ldots, 1.0\}. \tag{6}$$

Here, β weights the video endpoint in WiSE-FT, whereas α weights the fixed image-pretrained checkpoint in each Snapshot-Replay fusion.

Figure 7 reports the image and video averages under the same evaluation protocol as Table 5. WiSE-FT achieves its highest observed image score of 82.79 at $\beta = 0.5$, whereas its video score peaks at 61.77 at $\beta = 0.8$, where the image score is 82.25. Relative to $\beta = 0.5$, this improves the video average by 0.76 points while reducing the image average by 0.54 points. Snapshot-Replay obtains

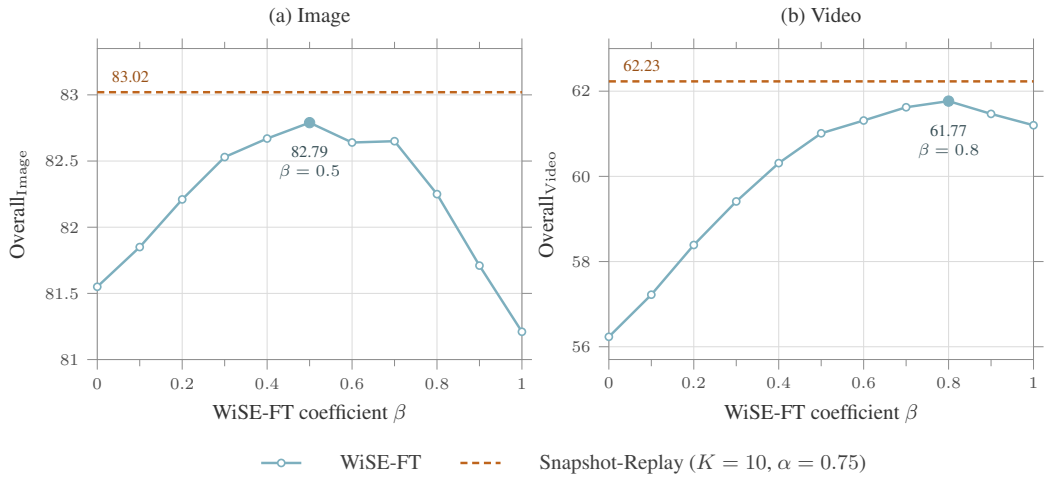


Figure 7: **WiSE-FT interpolation sweep.** The coefficient β controls the contribution of the continuous video fine-tuning endpoint, with $\beta = 0$ recovering the image-pretrained checkpoint and $\beta = 1$ recovering the video fine-tuning endpoint. Markers show the evaluated coefficients, and connecting segments serve as visual guides. Dashed lines show Snapshot-Replay at $K = 10$ and $\alpha = 0.75$. Image and video averages follow Table 5.

83.02 on images and 62.23 on videos, exceeding the respective WiSE-FT maxima by 0.23 and 0.46 points. It therefore outperforms every sampled coefficient on both averages.

Per-metric comparison. Table 20 compares Snapshot-Replay with the WiSE-FT coefficients yielding the highest image and video averages in the sweep. At the reported precision, Snapshot-Replay improves eight of ten image metrics and all nine video metrics over $\beta = 0.5$; ImageNet-R ties, while Flickr30K image-to-text retrieval decreases by 0.1 points. Against $\beta = 0.8$, it improves nine image metrics and five video metrics, with HMDB51 tied. The video gains include Kinetics-400, UCF101, MSR-VTT video-to-text retrieval, and both MSVD directions. However, both ActivityNet directions and MSR-VTT text-to-video retrieval favor $\beta = 0.8$. Thus, the average improvements extend across multiple benchmarks, while individual video retrieval metrics retain trade-offs.

D.5 Distillation Weight Sensitivity

We examine the classification–retrieval trade-off of pooled-patch distillation at $\lambda \in \{400, 1000, 4000\}$ against reference-token distillation at the default $\lambda = 4000$. All runs train for one pass over the same 400M-pair subset of COKE-1.7B with a global batch size of 32768. The learning rate warms up linearly to 10^{-3} over 1000 steps, then follows cosine decay to zero at 400M samples. Other settings follow §3.3.1, including four register tokens and eight additional reference tokens only in the reference-token configuration.

Table 21 shows that reducing the pooled-patch distillation weight from 4000 to 1000 raises the retrieval average from 68.33 to 70.30, while lowering the classification average from 72.77 to 70.33. Further reducing λ to 400 lowers classification to 67.65 without improving retrieval. Thus, weakening pooled-patch supervision improves retrieval at the expense of classification within the tested range. Reference-token distillation at $\lambda = 4000$ exceeds the best pooled-patch classification, retrieval, image, and video averages in this sweep by 1.38, 2.33, 2.55, and 1.44 points, respectively. These comparisons use the best pooled-patch weight for each aggregate separately. Reference-token distillation also outperforms every pooled-patch setting on all ten image and nine video metrics. The results show that reducing the pooled-patch distillation weight does not close the observed gap under the 400M-sample budget. This limited sweep complements the 1.6B-sample comparison in §3.3.1, without establishing superiority across all weights or training budgets.

Table 20: **Per-metric comparison with WiSE-FT.** The coefficients $\beta = 0.5$ and $\beta = 0.8$ yield the highest image and video averages, respectively, in the interpolation sweep. Classification reports accuracy and retrieval reports Recall@1. Bold marks the best reported score in each row, including ties. Overall averages follow Table 5.

	WiSE-FT		Snapshot-Replay
Metric	$\beta = 0.5$	$\beta = 0.8$	$K = 10,\ \alpha = 0.75$
Image classification			
ImageNet-1K	84.3	83.2	**84.4**
ImageNet-A	89.2	89.0	**89.6**
ImageNet-R	**96.5**	96.1	**96.5**
ImageNet-V2	78.7	77.9	**79.0**
ImageNet-Sketch	75.6	74.7	**75.7**
ObjectNet	84.2	82.8	**84.7**
Image-text retrieval			
Flickr30K T→I	85.8	85.7	**86.4**
Flickr30K I→T	**96.7**	**96.7**	96.6
COCO T→I	59.1	59.0	**59.4**
COCO I→T	77.8	77.4	**77.9**
$\text{Overall}_{\text{Image}}$	82.79	82.25	**83.02**
Video classification			
Kinetics-400	73.1	73.5	**74.1**
UCF101	88.5	87.6	**89.4**
HMDB51	56.7	**57.3**	**57.3**
Video-text retrieval			
ActivityNet V→T	45.7	**48.0**	47.4
ActivityNet T→V	48.6	**49.8**	49.1
MSR-VTT V→T	46.4	48.0	**49.1**
MSR-VTT T→V	49.6	**50.6**	50.2
MSVD V→T	81.8	82.4	**84.2**
MSVD T→V	58.7	58.7	**59.4**
$\text{Overall}_{\text{Video}}$	61.01	61.77	**62.23**

Table 21: **Distillation weight sensitivity at a** 400**M budget.** CLS and RET average six image classification and four image retrieval scores, respectively. Image averages all ten image metrics, and Video averages nine zero-shot video metrics.

	Pooled patch			Reference tokens
Metric	$\lambda = 400$	$\lambda = 1000$	$\lambda = 4000$	$\lambda = 4000$
CLS	67.65	70.33	72.77	**74.15**
RET	70.25	70.30	68.33	**72.63**
Image	68.69	70.32	70.99	**73.54**
Video	47.24	48.00	47.87	**49.44**

D.6 Scope of the Openness Claim

COKE uses public image-text and video-text corpora and released auxiliary models: Seed-X-PPO-7B-GPTQ-Int8 for translation, siglip2-so400m-patch16-256 for filtering, SigLIP 2-SO400M-384 for text initialization, and DINOv3-7B for visual supervision. Auxiliary pretraining data are excluded from COKE-1.7B's size. We plan to release COKE model weights, vocabularies, and evaluation code.