Acknowledgments

As part of their affiliation with UC Berkeley, the authors were supported in part by the National Science Foundation, the Ford Foundation, and/or the Berkeley Artificial Intelligence Research (BAIR) Industrial Alliance program. In addition, the work is supported by an Ai2 Young Investigator Award. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of any supporting entity.

Appendix

The appendix consists of the following further discussion:

- Appendix A discusses the model release.
- Appendix B discusses the datasets that we use for pre-training both TULIP and GeCo.
- Appendix C discusses the datasets that we use to evalaute TULIP.
- Appendix D discusses the implementation details for the generative data augmentation portion of our approach.
- Appendix G discusses some model detail configurations.
- Appendix E provides some additional experimental results complementing the main paper performance.
- Appendix F provides some visualizations of the self-attention weights of the TULIP model.

A. Code Release

For more information on the code, and for all models, see https://tulip-berkeley.github.io.

B. Training Data

We pre-train all models on the DataComp-1B dataset [23]. DataComp-1B is a large-scale dataset comprising approximately 1.4 billion image-text pairs, curated from the CommonPool collection of 12.8 billion samples. We also train with captions from Recap-DataComp-1B [32], a large-scale dataset where approximately 1.3 billion images from DataComp-1B have been re-captioned using LLaMA-3-powered LLaVA-1.5. The goal of this recaptioning process is to enhance the textual descriptions associated with web-crawled image-text pairs, addressing issues like misalignment, brevity, and lack of descriptive detail in original captions. The new dataset has longer and more diverse textual annotations, increasing from an average 10.22 words per caption to 49.43 words, capturing richer contextual details.

GeCo is fine-tuned on standard augmentations, as well as the following data:

WebVid-10M: WebVid-10M [3] is a large-scale video-text dataset designed to support video-language model training and text-to-video retrieval tasks. The dataset is automatically collected from the web using a pipeline similar to Conceptual Captions [44], ensuring diverse and naturally occurring video-caption pairs. A key feature of WebVid-10M is that it focuses on real-world, diverse, and multimodal video content, making it a more challenging and representative dataset compared to traditional manually annotated datasets. The dataset spans a wide range of video types, including people performing actions, nature scenes, travel vlogs, and instructional content. Unlike other large-scale video datasets such as HowTo100M, which rely on automated speech recognition (ASR) transcriptions (often introducing noise and weak supervision), WebVid-10M provides directly associated textual descriptions, resulting in higher-quality supervision for training vision-language models.

MVImgNet: MVImgNet [56] is a large-scale dataset of multi-view images, designed as a bridge between 2D and 3D vision by capturing real-world objects from multiple viewpoints. The dataset consists of 6.5 million frames extracted from 219,188 videos, covering 238 object classes with extensive annotations including object masks, camera parameters, and point clouds. Unlike single-image datasets like ImageNet, MVImgNet is built from videos, capturing objects from different angles, which naturally introduces 3D-aware visual signals.

C. Evaluation Datasets

ImageNet-1K: The ImageNet-1K dataset [15] is a large-scale benchmark dataset widely used for training and evaluating deep learning models in computer vision. It consists of approximately 1.28 million training images, 50,000 validation images, and 100,000 test images, categorized into 1,000 distinct object classes. These classes span a diverse range of objects, including animals, vehicles, tools, and everyday items, making it a comprehensive dataset for image classification tasks. ImageNet-V2 [43] is a re-evaluated version of the original ImageNet dataset, designed to assess the generalization ability of models trained on ImageNet-1K. It consists of 10,000 images curated using the same class distribution and data collection process as the original validation set but sourced independently to reduce potential dataset biases. ImageNet-ReaL [7] is a re-annotated version of the ImageNet validation set, created to provide more accurate and comprehensive labels. Unlike the original ImageNet-1K validation set, where each image is assigned a single ground truth label, ImageNet-ReaL introduces multi-label annotations, acknowledging that many images contain multiple valid object categories.

ObjectNet: ObjectNet [4] is a real-world test dataset designed to evaluate the robustness and generalization of image classification models beyond standard benchmarks like

ImageNet-1K. It consists of 50,000 images featuring objects from 313 categories, many of which overlap with ImageNet classes. Unlike ImageNet, ObjectNet introduces systematic variations in object orientation, background, and viewpoint, making it significantly more challenging for models.

iNaturalist-2018: iNaturalist-2018 [53] is a large-scale image classification dataset focused on fine-grained species recognition, designed to challenge models with real-world biodiversity data. It contains 437,513 training images and 24,426 validation images across 8,142 species, spanning diverse categories such as plants, insects, birds, mammals, and fungi. Unlike datasets like ImageNet, iNaturalist-2018 exhibits long-tailed class distributions, meaning some species have thousands of images while others have only a few, mimicking real-world imbalances in biodiversity data.

CIFAR-100: CIFAR-100 [31] is a small-scale image classification dataset designed for evaluating machine learning models, particularly in the context of deep learning. It consists of 60,000 color images of size 32×32 pixels, with 50,000 training images and 10,000 test images. The dataset contains 100 classes, each with 600 images, and these classes are further grouped into 20 superclasses (e.g., aquatic mammals, vehicles, flowers).

RxRx1: RxRx1 [47] is a biological image dataset designed for evaluating domain generalization in deep learning models, specifically in the context of cellular microscopy images. It consists of 125,510 images of human cells treated with various chemical perturbations, captured using high-throughput fluorescence microscopy. A key challenge in RxRx1 is that images come from multiple experimental batches across four cell types, introducing batch effects—systematic variations that can hinder model generalization.

fMoW: fMoW (Functional Map of the World) [14] is a large-scale remote sensing dataset designed to evaluate model performance on satellite image classification and change detection tasks. It contains over 1 million images from diverse geographic locations, covering 62 categories of functional land use and infrastructure, such as airports, military facilities, bridges, and solar farms. The dataset includes images captured under varied lighting conditions, seasonal changes, and resolutions, making it a challenging benchmark for real-world geospatial analysis.

Infographic: InfographicVQA [37] is a dataset designed for Visual Question Answering (VQA) on infographics, which are complex document images combining text, graphics, and data visualizations. The dataset consists of 5,485 images and 30,035 questions, with annotations requiring reasoning over various elements such as tables, figures, maps, and textual content. Unlike traditional VQA datasets, InfographicVQA places emphasis on elementary reasoning skills, including counting, sorting, and basic arithmetic operations.

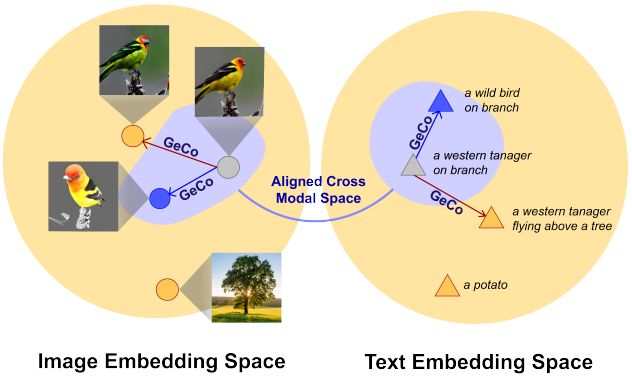

Figure C.1. (Top) GeCo generates positive (in blue region) and hard negative augmentations (in yellow region) of both images and text. Hard negative is closer to the 'positive region' while randomly sampled images or text are further.

Winnoground: Winoground [48] is a dataset introduced to evaluate the ability of vision-and-language models to perform visio-linguistic compositional reasoning. Each of the 400 examples in the dataset consists of two images and two captions, where both captions contain the same set of words arranged differently, leading to distinct meanings. The task requires models to correctly match each image with its corresponding caption, testing their understanding of how word order affects meaning in a visual context.

D. Data Augmentation

As discussed in subsection 3.2, we generate both positive view and negative views for contrastive learning. We show some example in Figure 5. To generate positive view of the image, we input positive embedding E_p. and a high image classifier free guidance (cfg) scale 5. To generate negative view of the image, we input negative embedding E_n to the model with a lower image cfg scale 3. To generate paraphrases for the image augmentation model, we use the prompt in Figure D.1, which can generate a positive and a negative example for an input caption. Figure C.1 gives additional insight into our data augmentation method.

E. Further Experimental Results

This section complements the experiments reported in the main paper by providing (i) comparisons against very recent baselines, (ii) evidence that TULIP learns transferable fine-grained visual representations, (iii) performance on the challenging UNED retrieval suite, and (iv) a more extensive ablation study. Unless otherwise stated, all results are

Given an input caption describing an image, generate two variants:

Positive Example: A paraphrased version that preserves the exact meaning using synonyms, grammatical reordering, or structural changes (e.g., active/passive voice).

Negative Example: A minimal, plausible alteration that subtly contradicts the original meaning. Prioritize compositional changes (e.g., swapped roles, spatial relations, object attributes, or verb actions) while keeping lexical overlap high. The negative should be visually distinct but textually similar to trick models.

Guidelines: Positive Paraphrase:
Use synonyms ("cube" → "square"), reorder clauses ("X beside Y" → "Y next to X"), or adjust syntax ("holding a leash" → "gripping a dog's lead").
Ensure no key details (objects, relationships, attributes) are altered.
Hard Negative:
Swap Roles/Relations: Invert subject-object relationships ("a man riding a horse" → "a horse beside a man").
Modify Prepositions/Spatial Logic: Change directional/positional cues ("left of" → "under").
Alter Attributes: Adjust colors, sizes, or quantities ("three red apples" → "two green apples").
Reorder Phrases with Identical Words: Use the same words in a different order to invert meaning ("plants surrounding a lightbulb" → "a lightbulb surrounding some plants").

Example: Input: "A chef in a white hat is slicing vegetables on a stainless steel counter while a cat watches from the windowsill."

Positive: "A cook wearing a white cap chops veggies on a shiny metal countertop as a feline observes from the window ledge." (Synonym substitution + rephrasing)

Negative: "A cat in a white hat is slicing vegetables on a stainless steel counter while a chef watches from the windowsill." (Role swap: "chef" ↔ "cat" + retained details create a contradictory but plausible scene.)

Figure D.1. The GeCo prompt.

obtained with the same hyper-parameters used in the main paper.

E.1. Comparison to Additional Baselines

Table E.2 benchmarks TULIP against **TIPS** [36], **SILC** [39], and the recently released **EVA-02 CLIP** [20] under matched model and resolution settings. TULIP clearly surpasses these strong competitors on Flickr, ImageNet and ObjectNet. On COCO, it matches the performance of the higher-resolution TIPS variant while using a lower input resolution.

E.2. Generality of the Learned Representations

To probe the low-level geometric understanding of TULIP, we follow El Banani et al. [19] and train a single linear layer on frozen features for **monocular depth estimation** (Table E.3) and **3-D correspondence** on NAVI (Table E.4). TULIP substantially narrows (and in most cases closes) the gap between vision-language pre-training and specialist self-supervised models, confirming that our contrastive *Generative-Contrastive* (GeCo) augmentations do not harm—and in fact improve—fine-grained spatial reasoning.

E.3. Retrieval on the UnED Benchmark

Table E.1 reports results on the recent UnED suite. TULIP establishes a new state-of-the-art on all seven domains and on the overall score, highlighting its robustness on long-tail, fine-grained retrieval tasks.

E.4. Extended Ablation Study

Table E.5 dissects the impact of every component added on top of the SigLIP backbone. The progressive gains confirm that each design choice—recaptioned data, intra-modal contrast, textual contrast, reconstruction loss, and finally the GeCo generative augmentations—contributes meaningfully and *additively* to the final performance. We caution, however, that carrying out ablations on dozens of datasets risks overfitting, so we restrict the study to a representative subset.

Table E.1. **UNED** retrieval (ViT-B). Numbers are Recall@1 / mMP@5.

Model	Cars		DF		GLDv2		iNat		Met		RP2K		SOP		**Overall**	
	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5
CLIP	76.4	69.5	42.8	18.3	15.9	9.6	42.3	32.3	22.6	9.9	57.9	39.4	54.1	29.4	49.6	27.8
MetaCLIP	84.4	78.9	54.3	23.6	15.4	9.4	46.9	36.5	17.0	8.2	56.1	37.5	62.2	36.1	56.8	33.2
SigLIP-2	94.6	93.1	58.7	25.7	14.2	8.7	48.1	38.2	32.1	14.7	73.1	55.3	65.8	39.4	61.1	36.9
TULIP	**96.8**	**95.8**	**72.6**	**32.3**	**22.2**	**12.7**	**58.9**	**49.1**	**54.1**	**22.8**	**72.5**	**55.8**	**68.0**	**41.2**	**67.0**	**40.9**

Table E.2. **TULIP vs. more baselines.** Res. denotes input resolution, FT indicates linear probing. Other numbers are zero-shot.

Model	Res.	COCO		Flickr					
		I→T	T→I	I→T	T→ I	IN	IN^{FT}	IN-V2	ObjNet
TIPS-g/14	224	73.3	58.1	93.4	82.1	79.6	86.3	—	—
TIPS-g/14	448	**74.0**	**59.2**	93.8	83.8	79.7	86.1	—	—
TULIP-g/16	384	73.0	57.8	**95.7**	**87.2**	**85.3**	**89.6**	80.0	88.6
EVA-02-CLIP-B/16	224	—	—	—	—	74.7	—	67.0	62.3
SILC-S-B/16	256	66.2	48.7	—	—	76.6	—	—	—
TULIP-B/16	224	**70.1**	**54.2**	93.9	81.8	**79.5**	85.9	**73.0**	**74.2**

Table E.3. Linear probing for **depth estimation**. Higher δ and lower RMSE are better.

Model	NYU				NAVI			
	δ_1	δ_2	δ_3	RMSE	δ_1	δ_2	δ_3	RMSE
CLIP	52.1	81.7	93.7	0.945	24.9	48.7	68.5	0.199
SigLIP	63.8	89.7	97.3	0.719	36.5	63.1	79.2	0.157
SigLIP-2	67.9	91.4	97.9	0.660	36.3	63.4	79.7	0.157
TULIP	**74.4**	**93.9**	**98.3**	**0.568**	**42.3**	**76.9**	**96.2**	**0.090**

Table E.4. **NAVI 3-D correspondence**. Retrieval at varying error thresholds (higher is better).

Model	R@0.01m	R@0.02m	R@0.05m	R@5px	R@25px	R@50px
SigLIP-2	12.0	26.4	62.5	1.61	9.21	22.0
TULIP	**14.4**	**29.1**	**64.8**	**1.87**	**12.2**	**24.9**

Table E.5. **Zero-shot ablations** (ViT-B). For COCO/Flickr we report I↔T retrieval.

Model	**Classification**					**COCO**	**Flickr**
	IN-val	IN-v2	IN-ReaL	ObjNet	IN-10s	T→I / I →T	T→I / I →T
SigLIP	76.2	69.5	82.8	70.7	69.9	47.2 / 64.5	77.9 / 89.6
+ Recaptioned	76.6	70.0	83.4	71.2	70.8	48.5 / 64.7	79.4 / 90.0
+ I/I	77.5	70.8	84.3	72.6	72.0	50.2 / 70.3	79.2 / 90.7
+ T/T	78.1	71.3	85.5	72.8	72.7	51.1 / 69.9	79.7 / 91.2
+ Reconstruction	78.5	71.7	85.3	73.4	72.5	52.1 / 69.3	80.0 / 92.8
+ GeCo (full TULIP)	**79.5**	**73.0**	**86.2**	**74.2**	**73.8**	**54.2 / 70.1**	**81.8 / 93.9**

Figure E.1. Visualization of the attention heads. Attention maps are averaged across transformer blocks, then up-sampled to the resolution of the original image.

Hyperparameter	ViT-G/16	ViT-SO400M	ViT-H-14	ViT-B-16
Embed Dim	1536	1152	1152	768
Init Logit Bias	-10	-10	-10	-10
Image Size	384	384	224	224
Patch Size	16	14	14	16
Layers (Vision)	43	27	32	12
Width (Vision)	1536768	1152768	1280	768
Head Width (Vision)	64	64	80	64
MLP Ratio	3.7362	3.7362	3.7362	4.0
Pooling	map	map	tok	map
Projection	none	none	linear	none
Context Length	70	70	70	70
Vocab Size	109871	109871	109871	109871
Tokenizer	tulip-tokenizer	tulip-tokenizer	tulip-tokenizer	tulip-tokenizer
Width (Text)	1152	1152	1024	768
Heads	16	16	16	12
Layers (Text)	27	27	24	12
No Causal Mask	True	True	True	True
Projection Bias	True	True	True	True
Pool Type	last	last	last	last
Norm Eps	10^{-6}	10^{-6}	10^{-6}	10^{-6}
Activation Approx.	tanh	tanh	tanh	-
Attentional Pool	False	False	False	False
Attn Pooler Queries	256	256	256	256
Attn Pooler Heads	8	8	8	8
Pos Embed Type	learnable	learnable	learnable	learnable
Final LN After Pool	False	False	False	False
Output Tokens	False	False	False	False
Timm Pool	map	map	avg	map
Timm Proj	none	none	linear	none
Timm Proj Bias	False	False	False	False
Timm Drop	0.0	0.0	0.0	0.0
Timm Drop Path	None	None	None	None

Table E.6. Comparison of Vision Transformer (ViT) Model Hyperparameters for different TULIP variants.

F. Attention Visualization

Figure E.1 shows a visualization of the attention heads of the So/14 model. We can see that similar to DINOv2, the model performs local semantic segmentation as an emergent behavior the in the attention weights.

G. Model Configurations

Table E.6 provides an overview of our model configurations, detailing key parameters such as image size, sequence length, hidden size, number of layers, and text context length. We follow SigLIP 2 to use So400M language encoder for ViT-G/16.