LR0.FM: LOW-RES BENCHMARK AND IMPROVING ROBUSTNESS FOR ZERO-SHOT CLASSIFICATION IN FOUNDATION MODELS

Priyank Pathak[1], **Shyam Marjit**[2], **Shruti Vyas**[1] **& Yogesh S Rawat**[1]
[1]University of Central Florida, [2]IIIT Guwahati
priyank@ucf.edu, shyam.marjit@iiitg.ac.in, {shruti, yogesh}@ucf.edu
https://ucf-crcv.github.io/lr0.fm

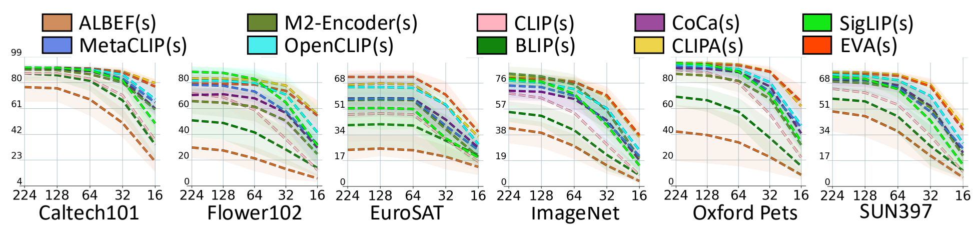

Figure 1: **Top-1 zero-shot classification accuracy (y-axis) vs resolution (x-axis)**: Backbones for foundation models are merged as shade, with average performance across backbones in the dark.

ABSTRACT

Visual-language foundation Models (FMs) exhibit remarkable zero-shot generalization across diverse tasks, largely attributed to extensive pre-training on large-scale datasets. However, their robustness on low-resolution/pixelated (LR) images, a common challenge in real-world scenarios, remains underexplored. We introduce **LR0.FM**, a comprehensive benchmark evaluating the impact of low resolution on the zero-shot classification performance of *10* FM(s) across *66* backbones and *15* datasets. We propose a novel metric, **Weighted Aggregated Robustness**, to address the limitations of existing metrics and better evaluate model performance across resolutions and datasets. Our key findings show that: (i) model size positively correlates with robustness to resolution degradation, (ii) pre-training dataset quality is more important than its size, and (iii) fine-tuned and higher resolution models are less robust against LR. Our analysis further reveals that the model makes semantically reasonable predictions at LR, and the lack of fine-grained details in input adversely impacts the model's initial layers more than the deeper layers. We use these insights and introduce a simple strategy, **LR-TK0**, to enhance the robustness of models without compromising their pre-trained weights. We demonstrate the effectiveness of **LR-TK0** for robustness against low-resolution across several datasets and its generalization capability across backbones and other approaches. *Code is available at this link.*

1 INTRODUCTION

Vision-Language Foundation Models (FMs), such as CLIP (Radford et al., 2021), LLaMA (Touvron et al., 2023), and other variants, have shown extraordinary generalization capabilities across a wide range of downstream tasks, including image classification (Ilharco et al., 2021), object detection (Zhong et al., 2022), and semantic segmentation (Xu et al., 2023). These models benefit from large-scale, multi-modal pre-training on diverse datasets like DataComp-1B (Gadre et al., 2023) and LAION-5B (Schuhmann et al., 2022), enabling them with zero-shot capabilities. Although these models excel on high-resolution benchmarks, their performance with low-resolution (LR) pixelated images, a common real-world challenge, remains adequately underexplored.

Low-resolution images frequently arise in various practical scenarios, such as surveillance footage (Davila et al., 2023), satellite imagery (Patil et al., 2017), and privacy-protected pixelated

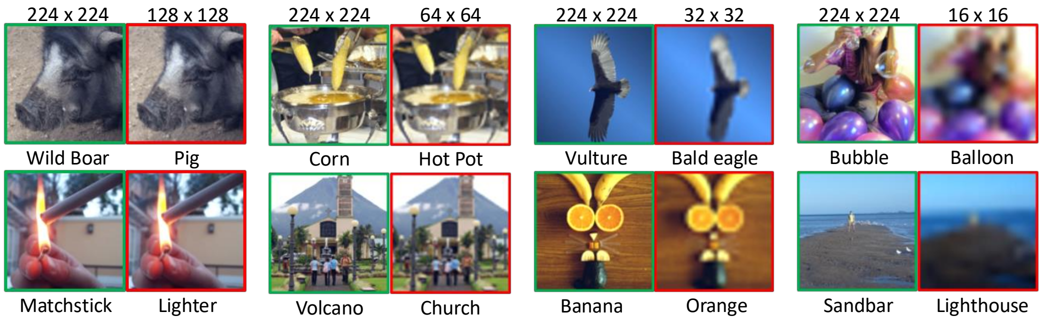

Figure 2: **Zero-Shot misclassifications:** EVA-CLIP [2023a] correct classification at 224×224 (green) & misclassification at lower resolution (red). However, ImageNet labels-based mispredictions are semantically reasonable (humans), indicating viability of pre-trained weights at low resolution.

data (Zhou et al., 2020) *etc*. In these cases, details crucial for accurate classification may be obscured by artifacts like pixelation and compression, leading to substantial performance degradation. For instance, small objects (faces) within larger images (Cheng et al., 2019) pose unique challenges, often requiring models to rely on limited visual cues. Given the widespread presence of LR images in real-world applications, it is crucial to understand how robust FMs are in these settings.

Motivated by this, we present an in-depth benchmarking study of FMs, focusing on their zero-shot classification performance under LR conditions. We introduce LR0.FM, a comprehensive benchmark that evaluates **10 foundation models** across **66 backbones** and **15** diverse image classification **datasets**, ranging from large-scale datasets like ImageNet (Deng et al., 2009) to fine-grained and texture-specific datasets like Oxford Pets (Parkhi et al., 2012) and DTD (Cimpoi et al., 2014). Our study systematically examines the effects of resolution degradation, revealing key insights into how model size, pre-training dataset quality, and fine-tuning impact robustness in LR scenarios.

Metrics for measuring robustness γ (Schiappa et al. (2024b)) and its averaging across datasets (SAR) have some limitations; 1) They can produce misleadingly high scores when models perform poorly on challenging datasets, and 2) They tend to ignore certain datasets, skewing the overall comparison. To address these, we propose a new metric, **Weighted Aggregated Robustness (WAR)**, which provides a more balanced evaluation by considering performance drops across datasets more fairly.

Our analysis reveals several interesting insights. Larger models tend to maintain robustness better when faced with LR inputs, while the quality of the pre-training dataset is more crucial than its size in preserving performance. Furthermore, fine-tuned models and those with higher-resolution inputs significantly underperform against resolution drop. We also observe that although models struggle at low-resolution (fig. 1) and loss of fine-grained details (fig. 2: *e.g.* Vulture vs Bald Eagle, Bubble vs Balloon *etc.*), their predictions often remain semantically reasonable, even at extreme resolutions (fig. 2: *e.g.* Orange vs Banana, Church vs Volcano *etc.*). *Supplementary* demonstrates more examples (including real-world) where such mispredictions are made. This suggests a solution for low-resolution does not require extensive modifications to the model and its pre-trained weights.

Based on these insights, we propose a simple yet effective solution, **LR-TK0: LR-Zero-Shot Tokens**, which introduces low-resolution-specific tokens to enhance robustness without altering the pre-trained model weights. Our method preserves the model's semantic reasoning capabilities while compensating for the loss of fine-grained detail, offering a feature super-resolution-like approach (Chen et al., 2024). By training on synthetic diffusion-based high-resolution images, LR-TK0 improves performance in low-resolution zero-shot classification tasks, making FMs more robust for practical, real-world applications.

In summary, we make the following contributions in this work,

1. We present ***LR0.FM***, a comprehensive benchmarking of Visual-Language Foundation Models (FMs) on zero-shot classification of low-resolution images, providing several key insights. To the best of our knowledge, no prior work has explored this aspect of FMs.
2. We introduce a simple and effective method, ***LR-TK0***, to enhance model robustness against low-resolution inputs without altering the pre-trained weights.
3. We introduce Weighted Aggregated Robustness (**WAR**), a novel robustness metric for evaluating models under challenging conditions, overcoming the limitations of existing metrics.

Table 1: **Benchmark Models (66 Backbones):** Pre-training is image-text pairs from datasets like DataComp-1B (DC-1B) (Gadre et al., 2023), Conceptual Captions (CC) (Sharma et al., 2018), Conceptual 12M (C-12M) (Changpinyo et al., 2021). Text Encoders are mostly modified vanilla transformers (Tran.)(Vaswani et al., 2017). Vision backbones use (modified) ViTs (Dosovitskiy, 2021).

Models	#Backbones	Pre-training (Dataset / Size Billion:B & Million:M)		Text Encoder
CLIP [2021]	4 ViTs & 5 ResNets	WIT-400M [2021]	400M	Tran. [2019]
OpenCLIP [2021]	8 ViTs	DC-1B, LAION-2B[2022], DFN-5B[2023]	1B-5B	Tran.[2021]
MetaCLIP [2024a]	8 ViTs	*Self*	400M-2.5B	OpenCLIP
CLIPA (v1&v2) [2023b; 2023c]	7 ViTs	DC-1B, LAION-2B [2022]	1B-2B	Autoregressive Tran. [2017]
SigLIP [2023]	8 ViTs	WebLI [2023b]	10B	Tran.
CoCa [2022]	3 ViTs	LAION-2B [2022], COCO [2014]	2B	Tran. Decoder
M^2-Encoder[2024]	3M^2-Encoder	BM-6B [2024]	6B	Magneto [2023]
ALBEF [2021]	4 ALBEF (ViT)	COCO [2014], Visual Genome [2017], CC, SBU Captions [2011], C-12M	4M-14M	BERT [2019]
BLIP [2022b]	8 ViTs	ALBEF [2021], LAION-400M [2021]	14M-129M	BERT [2019]
EVA-CLIP(&18B) [2023a; 2023b]	8 EVA(s) (ViT(s))	LAION-400M[2021], LAION-2B[2022], Merged-2B [2023a]	400M-2B	OpenCLIP

2 Related Works

Foundation Models (FM): Large-scale models (Kirillov et al., 2023; Girdhar et al., 2023), pre-trained on massive datasets, demonstrate generalization across numerous downstream tasks. For example, CLIP (Radford et al., 2021) embeds $\sim$400 million image-text pairs in a shared feature space for zero-shot image classification and image-text retrieval. It is also effective in other domains like video-text retrieval (Luo et al., 2022), and video and audio understanding (Lin et al., 2022; Guzhov et al., 2022). Joint vision-text learning has also succeeded in tasks such as self-supervision (Miech et al., 2020), few-shot (Alayrac et al., 2022), multi-modal retrieval (Yu et al., 2022) *etc.* However, the robustness of these models against *real-world challenges e.g.* harmful images (Qu et al., 2024), image quality (Wu et al., 2024), text quality (Xu et al., 2024b), *etc.* requires further exploration.

Zero Shot: Zero Shot/Open-set/In-the-wild image classification predicts an unseen class by matching the image with labels (Sun et al., 2023a). In the past, traditional models have been tested for their zero-shot capabilities (Chao et al., 2016; Xian et al., 2017), however, FMs are better suited for this task. Benchmarking their zero-shot capabilities is a relatively newer area of research (Schiappa et al., 2024a; Schulter et al., 2023). To assess the performance comprehensively, we have expanded the pool of models from traditional 10-11 FM backbones *e.g.* 4 backbones (Li et al., 2022a), 9 backbones (Liu et al., 2024), 6 backbones ((Zhang et al., 2024)) *etc.* to *66 backbones*.

Low Resolution (LR): LR images are captured in various practical scenarios and are sometimes used intentionally for computational cost reduction (RECLIP (Li et al., 2023a)). LR benchmarks mostly focus on face recognition (Luevano et al., 2021; Li et al., 2018; Tirupattur et al., 2021), with some work in zero-shot/unconstrained recognition (Li et al., 2019; Cheng et al., 2019). Super Resolution (Ohtani et al., 2024; Gao et al., 2023) are often domain-specific or restores only $\geq 64\times64$. However, there is a lack of study on the robustness of FM(s) against real-world challenges (Xu et al., 2024b; Schiappa et al., 2023), with no previous work on very LR. We benchmark FM(s) against LR images and propose a lightweight solution for improving robustness, without training on any of the target datasets (Chen et al., 2024).

3 Benchmarking Setup

Model: Table 1 lists all 10 Foundation models used in our benchmarking[1]. CLIP, OpenCLIP, MetaCLIP, CLIPA, and SigLIP use the same ViT model with different pre-training datasets and slight architectural modifications (*e.g.* layer norm position, token masking *etc.*). M^2-Encoder (built on top of CoCa), ALBEF, and BLIP use modified cross attention between text and vision transformers. EVA-CLIP is a family of models equipped with recent advancements *e.g.* architectural modifica-

[1]EVA-CLIP (Sun et al., 2023a) & EVA-CLIP-18B (Sun et al., 2023b) merged into one.

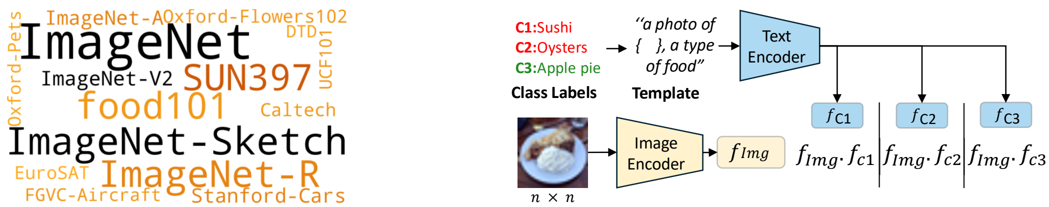

Figure 3: *Left*: **Dataset**: Size $\propto$ log # test images, and color gradient $\propto$ # of test classes **orange** is 10 & **black** is 1000 classes). *Right*: **Zero Shot Evaluation**: Food-101 image (32×32) generates image embeddings f_{Img}, while class labels are filled in templates (1 shown) generating text embeddings (averaged across templates). The dot product of f_{Img} with text features gives classification logits.

tions, token dropping, training via distillation *etc.* surpassing all existing works. Backbones are referred to using their publicly available pre-trained weights, *e.g.* CLIP-ViT L (400M), which means: CLIP model ViT-L architecture, pre-trained on 400 million datasets. 'B' would indicate a billion.

Dataset: Figure 3 (left) highlights benchmarking datasets size and the number of classes for: Imagenet [2009], ImageNet-A [2021b], ImageNet-V2 [2019], ImageNet-R [2021a], ImageNet-Sketch (ImageNet-SK) [2019], Caltech101 [2007], DTD split-1 (DTD) [2014], Food101 [2014], SUN397 [2014] Stanford Cars (Cars) [2020], FGVC Aircraft (Aircraft) [2013], Oxford Pets (Pets) [2012], Oxford Flowers102 (Flowers102) [2016], EuroSAT [2019], UCF101 [2012]. Details in *Supplementary*.

Zero-Shot Image Classification We adopt CLIP (Radford et al., 2021) evaluation protocol for all the models as shown in fig. 3 (right). Image encoder generates embeddings for images, while test labels are used with dataset-specific templates (multiple templates, *Supplementary*) *e.g.* "a photo of a [label]". Model's Text encoder generates final text embeddings (averaged across all templates) for the class label. The dot product of visual and text embeddings produces class logits, with the highest logit score determining the predicted class. Accuracy is computed using Top-1 match.

Low Resolution: Models are evaluated on their pre-trained resolution, namely 224×224 256×256, 378×378 *etc.* Low resolution is simulated by downsampling HR images to 16×16, 32×32, 64×64, and 128×128 using bicubic interpolation, followed by model specific preprocessing similar to their HR counterparts, *e.g.* resizing to 224×224, center crop, *etc.* Performance degradation starts below 64×64 (fig. 1), so we focus mainly on 16×16 and 32×32. This downsampling mimics pixelation as seen in low-resolution cameras (*e.g.* self-driving cars) and distant images (*e.g.* CCTV), *etc.*

Evaluation Metrics: We represent top-1 accuracy on the dataset 'D' with a resolution $n \times n$ as $A_n^D \in [0, 1]$, *e.g.* HR accuracy $A_{HR}^D \geq A_n^D$ (LR accuracy), where HR is model specific $\in$ {224, 256, 372, 384, 512}. Top-1 scores averaged across datasets is **ACC-n**. Robustness against artifacts (Schiappa et al., 2024b) is measured by relative robustness ($\gamma_n^D = 1 - (A_{HR}^D - A_n^D)/A_{HR}^D$). γ_n^D is dataset-specific, and it is common to average scores across datasets for model comparison, denoted by **Simple Aggregated Robustness (SAR-n)**. *Higher number indicates more robustness*. However, there are two significant issues with γ_n^D and SAR-n:

Problem A) Misleading high robustness: If the model performs poorly on a challenging dataset *i.e.* performance close to random predictions, then downsampling will likely maintain this random prediction with minimal drop in accuracy, giving abnormally high robustness score. *Ex.* 'ALBEF (4M)' for Aircraft dataset, ($A_{\text{rand}}^{\text{aircraft}} = 1\%$), $A_{HR}^{\text{aircraft}} = 2.7\%$, $A_{16}^{\text{aircraft}} = 1\%$, $\gamma_{16}^{\text{aircraft}} = 37\%$, *i.e.* random predictions yields$\sim$40% robustness (40% robustness is among the highest, more in *Supplementary*).

Solution: Improved Relative Robustness Γ_n^D**:** A naive solution is to calculate relative robustness only for correct predictions at the HR resolution. However, tracking predictions for each model across all datasets might not be scalable, especially if the dataset contains millions of images. We propose *zero-ing out robustness near random predictions*. We first define *accuracy gap* for the model on a dataset with 'C' classes as $\mathcal{E}_D = A_{HR}^D - A_{rand}^D$, with $\mathcal{E}_D \in [0, 1]$, and $A_{rand}^D = 1/C$ represents random prediction accuracy[2]. If $A_{HR}^D >> A_{rand}^D$, $\mathcal{E}_D$ will be high. Conversely, if $A_{HR}^D \simeq$ random prediction, $\mathcal{E}_D \to 0$. Using $\mathcal{E}_D$, we compute **improved relative robustness** Γ_n^D as

$$\Gamma_n^D = \gamma_n^D \times (1 - e^{-\alpha(\mathcal{E}_D)^2}) \quad | \quad \alpha >> 1 \quad \& \quad 0 \leq \mathcal{E}_D \leq 1 \tag{1}$$

[2]Random guessing one of the 'C' class yields $1/C$ accuracy, referred to as A_{rand}^D in this work.

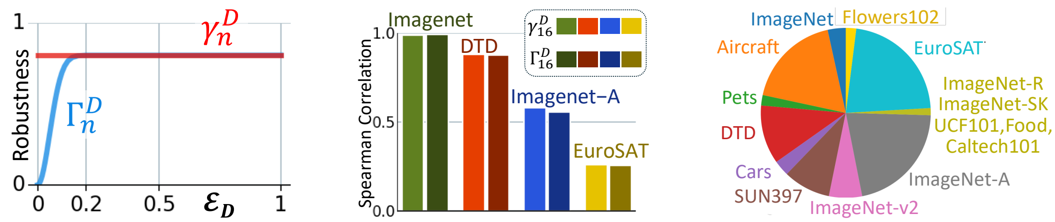

Figure 4: *Left*: **Improved Γ_n^D vs traditional γ_n^D**: $\Gamma_n^D \approx \gamma_n^D$ except near random predictions ($\mathcal{E}_D \rightarrow 0$). *Mid*: **Correlation** between the ordering of models after averaging of robustness (SAR) across datasets (γ_{16}^D & Γ_{16}^D) with dataset's true ordering. SAR final ranking ignores datasets like EuroSAT (0.26). *Right*: **Optimized dataset weights** for WAR-16. *Supplementary* contains numeric value.

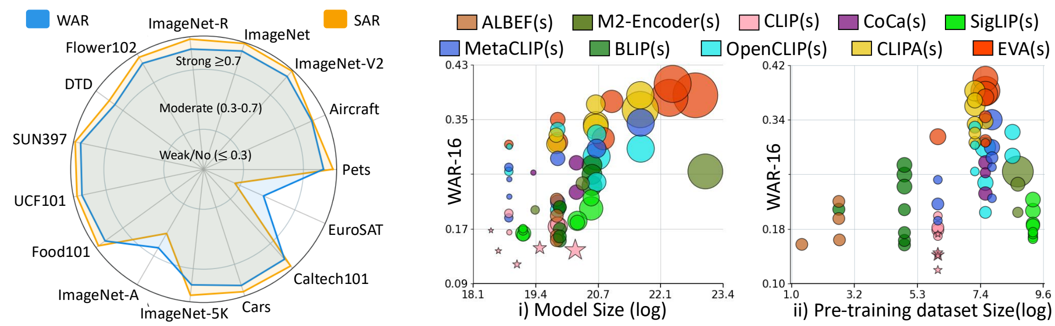

Figure 5: Evaluations at 16×16. *Left:* **SAR vs WAR:** WAR improves the correlation (between the ordering of models after aggregation with individual datasets) for EuroSAT (0.26 → 0.49 and ImageNet-A (0.56 → 0.68), both computed via Γ_{16}^D. *Right*: **i) Model Size & ii) Pre-training dataset size** positively impacts robustness. (i) Dot size ∝ GFLOPs, no impact on robustness (ii) Dot size ∝ Model Size, positively impact robustness. ResNets (⋆), and transformers (◯).

when $\mathcal{E}_D \sim 0$ *i.e.* near random predictions, $\Gamma_n^D \sim 0$, otherwise $\Gamma_n^D \approx \gamma_n^D$, as shown in fig. 4 (left). Hyperparameter α is the rate at which Γ_n^D declines as accuracy approaches random prediction. We chose $\alpha = 200$ as a middle between 100 (the drop at $\mathcal{E}_D \sim 0.2$) and 500 (the drop at $\mathcal{E}_D \sim 0$).

Problem B) SAR overlooks datasets: When comparing models, their robustness scores are averaged across datasets (SAR). Ideally, model ranking, after averaging robustness across datasets, should stay consistent with their rankings on individual datasets. However, fig. 4 (mid) shows the rankings of 66 models after averaging correlate (Spearman Rank) highly with ImageNet (0.99) and DTD (0.88), but only moderately with ImageNet-A (0.56) and weakly / not with EuroSAT (0.26). Most datasets follow the ImageNet trend, influencing the final model rankings and minimizing the impact of datasets like ImageNet-A and EuroSAT (behave differently) as if these datasets aren't present.

Solution: Weighted Aggregated Robustness: Averaging the robustness scores gives each dataset score of 1. We propose adjusting the dataset weights so that the *model rankings after aggregation reflect each dataset fairly* (fig. 4 (right)). Weights are optimized such that the correlation (Spearman) between the model rankings after the weighted average and individual dataset rankings are maximized. The weighted sum of robustness is: **WAR-n**$= \sum_d^{\text{Datasets}} |\Gamma_n^d \times w_n^d| / \sum_d^{\text{Datasets}} |w_n^d|$, where w_n^d is dataset weight, and Γ_n^d is dataset-specific improved robustness score for the resolution $n \times n$.

We use Ax tool (Bakshy et al., 2018) for optimizing the weights of the dataset $w_{16}^d \in [0.1, 1]$ such that the Spearman correlation (SC) between the final model ranking obtained after the weighted averaging and individual dataset ranking is maximized on empirically found (more in *Supplementary*):

$$0.95 \times \big(\text{SC(Imagenet)} + \text{SC(ImageNet-V2)} + \text{SC(DTD)}\big) + \text{SC(ImageNet-A)} + \text{SC(EuroSAT)} \quad (2)$$

Optimizing w_{16}^d may give minimal weights to some datasets, thus WAR-n may not reflect the true robustness and is more apt for model comparisons, representing all the datasets. Hence we use both Weighted Aggregated Robustness (WAR) using improved relative robustness Γ_n^D (eq. (1)) and simple averaging (SAR) using traditional robustness γ_n^D for evaluating models. *Note*, γ_n^D **and** Γ_n^D **measure dataset robustness while SAR and WAR measure averaged robustness across the datasets**.

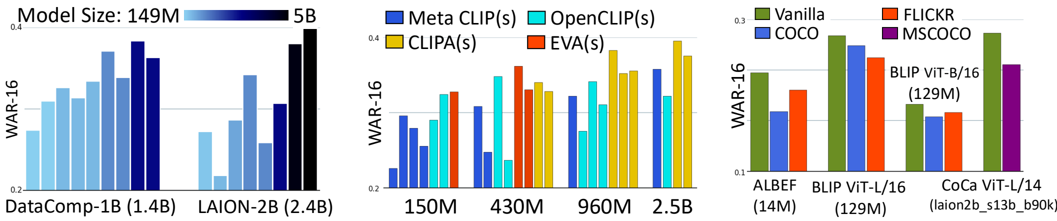

Figure 6: *Left:* **DataComp-1B vs LAION-2B**: Smaller DataComp-1B pre-training helps robustness. Models are ordered via size. *Mid:* **Model comparison w/o Size**: Models binned into size buckets ($\pm$30M). *Right:* **Fine-tuning degrades** robustness. (*left* & *mid*): bigger models are more robust.

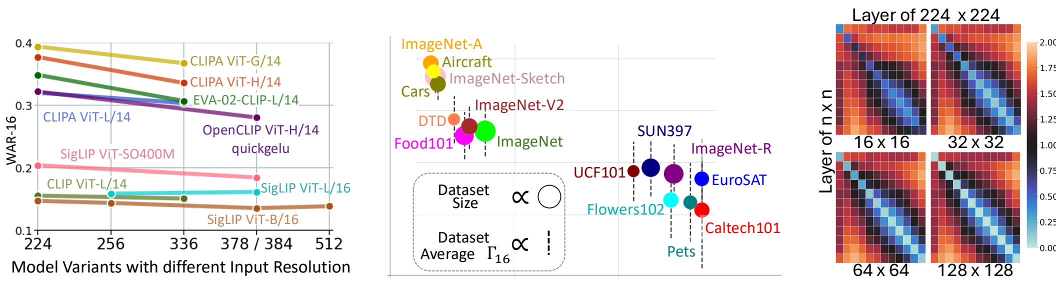

Figure 7: *Left*: **High Input Resolution Model** are less robust. *Mid*: **t-SNE of Dataset robustness** Dataset represented via 66 models robustness (Γ_{16}^{D}), indicates 3 clusters. *Right*: **Layers-wise features L2 similarity**: $n \times n$ model layers similarity w/ 224×224 ones, for EVA02-B-16. For a given heatmap (*e.g.* 16×16), the lower right indicates the similarity of deeper layers (brighter means more similar), while the upper left represents non-similar shallow layers (dull means less similar).

4 Benchmarking Analysis

Proposed WAR Metrics: Spearman correlation between the rankings of 66 models, calculated using SAR and WAR averaging of relative robustness Γ_{16}^{D} (across all datasets), and the individual dataset rankings is shown in fig. 5(left). WAR shows a slight decrease in avg. correlation (SAR-16 *0.89* vs WAR-16 *0.87*), but it also improves the representation of EuroSAT & ImageNet-A. The correlation score for EuroSAT increased from a weak/no correlation of 0.26 to a moderate 0.49.

Model Architecture / Pretraining: Figure 5 (right, (i)) shows, on average, **larger model** (x-axis) are **more robust**. Among the models, CLIP-ResNets (stars) are the least robust (compared to transformers (dots)) while EVA, MetaCLIP, CLIPA, and OpenCLIP exhibit the highest robustness against the LR. Higher GFLOP (size of dots) weakly impacts robustness with too many exceptions.

Pretraining 'Quality over Quantity': Figure 5 (right (ii)) shows pre-training dataset size weakly correlates with robustness, with exceptions like SigLIP (10B), and M2-encoder (6B) performing worse. Models pre-trained on DataComp-1B generally outperform those pre-trained on LAION-2B, despite having over 500M fewer image-text pairs (fig. 6 (left)). This suggests that the **model** and **quality of pre-training** have a greater impact on robustness **than the quantity of pre-training**.

Model Specific: We remove architectural size advantages by categorizing top-performing models into parameter buckets as shown in fig. 6 (mid). For smaller models (150M and 430M parameters), OpenCLIP matches EVA and outperforms MetaCLIP and CLIPA, despite these two being built on top of OpenCLIP. However, for larger models, this trend reverses, with EVA-CLIP remaining superior for comparable sizes. Two factors contribute to performance discrepancies within models of the same parameter size: **(1) Fine-tuning:** ALBEF and BLIP fine-tuned variants are less robust on EuroSAT and Aircraft, reducing their overall robustness (fig. 6 (right)) **(2) Higher input resolution:** Models with higher input resolutions (*e.g.* 336×336) are generally less robust than their 224×224 counterparts, likely due to increased interpolation from 16×16 to higher resolutions (fig. 7 (left)).

Dataset Specific: Relative robustness of 66 models on each dataset forms its robustness vector representations. Representing these vectors using t-SNE (fig. 7 (mid)), reveal three major dataset clusters: high-robustness (long bars) (*e.g.* Caltech101), weakly robust (medium bars) (*e.g.* ImageNet), and least robust (smallest bar) (*e.g.* , ImageNet-A). This indicates that **low-resolution performance varies by dataset**, which warrants a deeper dive into dataset-specific robustness, *left as future work*.

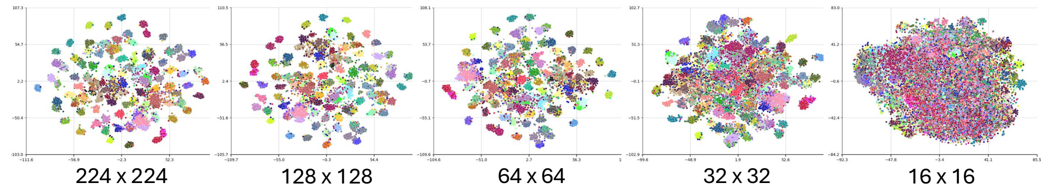

Figure 8: **Feats t-SNE**: EVA-02-CLIP-B/16 test features for Food-101, colored using class labels. With low resolutions (16×16, and 32×32), features become indistinguishable, thereby overlapping.

Figure 9: **Super resolution at 16×16**: Image from Pets (*left*) and Food102 (*right*). Models include AddSR [2024], BSRGAN [2021], ESRGAN [2018], IDM [2023], Inf-DiT [2024], and Swinir [2021].

Inside Model: Figure 1 shows the accuracy of all models first drops at 64×64, with a more significant decline after 32×32. EVA-B/16 features t-SNE (fig. 8) shows **features become indistinguishable as resolution decreases**. Inside the model, Figure 7 (right) shows the pairwise similarity (L2 distance) (Kornblith et al., 2019) between layers of models trained at different resolutions with the 224×224. Diagonal elements (i^{th} layer of $n\times n$ model similarity with i^{th} layer of a model trained at 224×224), is **more similar towards the deeper end** (lower right, the similarity is brighter), **than the initial layers** (upper left, the similarity is dull). Additionally, model similarity increases with resolution, while layers remain differentiable at all resolutions (dull non-diagonal values).

5 Proposed Method: LR-TK0

Figure 2 reveals two key insights: i) LR lacks fine-grained details ii) FM(s) make semantically reasonable predictions even at 16×16, highlighting the importance of preserving semantic capabilities (pre-training). While super-resolution (SR) methods could restore lost details without affecting models, zero-shot SR for very low resolutions ($\leq 64\times64$), doesn't work well in practice, as shown in fig. 9, where SR models fail to reconstruct out-of-domain images at 16×16. To enhance model robustness against low resolution, our solution **LR-TK0** adds trainable LR tokens on top of frozen transformers (preserving the pre-trained weights). These LR tokens learn to bridge the gap between the high-resolution (HR) and low-resolution (LR) domains, via self-supervised distillation (section 5.1). We train these tokens on synthetically generated diffusion-based images (Section 5.2) in a task-agnostic setting, ensuring the model is not exposed to any of the 15 target datasets.

5.1 LR tokens

To preserve the zero-shot capabilities of the model; *pre-trained weights of the model are frozen*. Instead, additional trainable tokens, referred to as "LR Tokens", are added on top of the spatial tokens after RGB to patch tokens conversion (patchification) and before each transformer block. As shown in fig. 10 (left) # LR tokens = # Spatial tokens × (N+1) blocks. These tokens aim to compensate for the loss of details in low resolution, thereby enhancing the model's interpretability of LR images. Contrary to prompt learning (Jia et al., 2022), where task-specific tokens are concatenated to the spatial tokens, ours are added/merged. Figure 7 (right) indicates LR feature at the initial layer deviates more than the later ones, thus LR tokens are added at every block.

LR-TK0 Technique: We adopt the multi-scale paradigm (Chen et al., 2019) *i.e.* training multiple low resolutions per HR image, given its success in the LR domain. Model without LR tokens (frozen pre-trained weights) acts as a teacher generating feature representations for HR images, as true embedding $f^{\mathbf{T}}_{HR}$. In contrast, LR tokens (& pre-trained model) act as student, generating

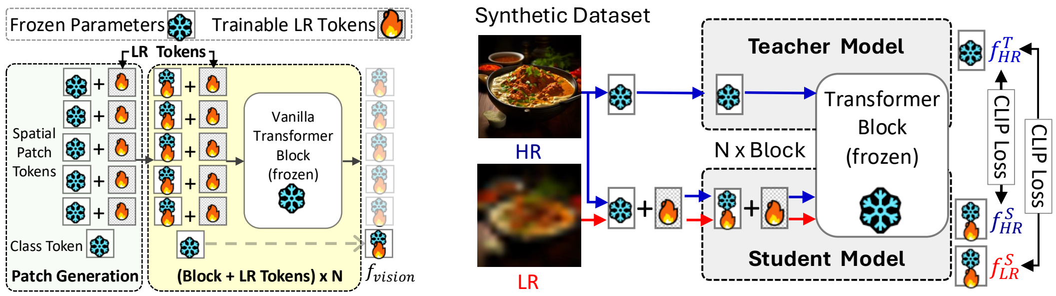

Figure 10: Fire (& ice) icons represent trainable (& frozen) parameters. *Left:* **LR tokens** are added to the frozen spatial patches (white) after patch generation, before each frozen transformer block, and class token as a final feature. *Right:* **LR-TK0:** Multi-scale training (only 1 shown for simplicity). Teacher (w/o LR tokens) generates f^T_{HR} (HR), Student (w/ LR tokens) generates both f^S_{HR}, f^S_{LR}.

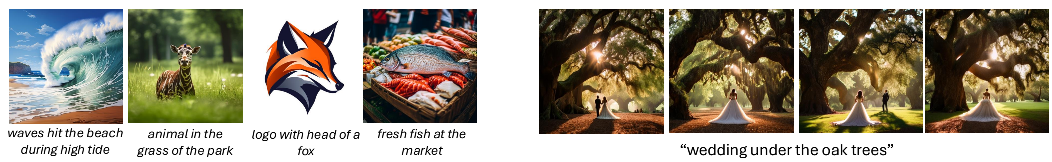

Figure 11: **Synthetic Images**: (*Left*) Images generated using PIXART-α [2023a] using randomly sampled captions from Conceptual Captions [2018]. (*Right*) Multiple images per caption.

embeddings for both HR ($f^{\mathbf{S}}_{HR}$) and LR image(s) ($f^{\mathbf{S}}_{LR}$) as shown in fig. 10 (right). $f^{\mathbf{S}}_{HR}, f^{\mathbf{S}}_{LR}(s)$ are matched with $f^{\mathbf{T}}_{HR}$ using a contrastive loss (Radford et al., 2021), similar to text and image alignment. Anchoring HR-LR features around frozen teacher avoids direct matching of HR-LR embeddings, preventing pulling the HR features towards LR ones (converging into one) (Khalid et al., 2020). This also ensures features w/ and w/o spatial tokens remain similar (regularization). Feature matching doesn't require any labels for these synthetic images, aka **unsupervised**. It also **task agnostic**, *i.e.* doesn't involve any model task-related characteristics (classification in this case).

5.2 Synthetic HR Dataset

We use the diffusion model PIXART-α (Chen et al., 2023a) to generate synthetic HR images, via 7,000 randomly sampled captions from Conceptual Captions (Sharma et al., 2018). We expand our training set by creating multiple images (variations, human observation) per caption as shown in fig. 11. Conceptual Captions are commonly used in pretraining many zero-shot models (table 1), and using synthetic diffusion-based images helps LR tokens capture a wide range of domains, ensuring generalized training. **Random captions avoid targeting any specific dataset**. To our knowledge, our work is the first to train a model on synthetic diffusion images for zero-shot evaluation, contrary to training on a subset of target datasets (Chen et al., 2024). Following the multi-scale paradigm, we downsample HR images to a randomly sampled spatial resolution (height = width) from three LR resolution buckets [16,32], [32, 64], [64, 128], forming HR-LR image pairs.

Zero-Shot: If 7,000 (or fewer) concepts/captions can consistently enhance model performance across 15 datasets, it suggests that the model is likely learning the relationship between HR and LR features rather than exploiting shortcuts. This is supported by greater improvements at LR (16×16) compared to HR (128×128). If the model somehow cheats the zero-shot evaluation using diffusion-generated images, we would expect similar or better performance improvements at HRs.

6 Proposed Method: Experimentation & Ablation

Implementation Details Models are trained with 7K captions (& 30 images/captions) in a multi-scale paradigm. EVA is trained for 200 epochs, while MetaCLIP and OpenCLIP are for 10 epochs. Evaluation metrics (section 3): SAR (simple averaging of γ^D_n), WAR (weighted averaging of Γ^D_n), and Acc (average top-1). Higher number means better performance. Vanilla model's HR accuracy computes the accuracy gap $\mathcal{E}_D$, and dataset weights derived for 16×16 used for all resolutions (more in *Supplementary*). **'EVA-02-CLIP-B/16' (EVA-B/16)**, is used for all our model-level analysis.

Table 2: **LR-TK0 improvement on Foundation models:** 'Meta-B/16': MetaCLIP-ViT-B/16 (2.5B), 'OC-B/16': OpenCLIP-ViT-B/16. Higher number $\propto$ better performance.

Model	# Param	16×16			32×32			64×64			128×128			224×224		
		SAR	WAR	Acc	SAR	WAR	Acc	SAR	WAR	Acc	SAR	WAR	Acc	SAR	WAR	Acc
EVA-B/16	149.7M	38.0	30.7	28.1	74.4	64.8	53.5	92.4	85.8	65.2	98.4	96.1	68.8	100	100	69.6
+LR-TK0	155.2M	42.4	35.4	31.3	75.3	66.4	54.1	91.8	85.9	64.8	97.8	95.5	68.3	99.1	98.7	69.0
Meta-B/16	149.6M	32.1	27.2	23.4	65.3	54.4	47.0	89.5	83.6	62.9	98.5	96.7	68.5	100	100.0	69.4
+LR-TK0	151.6M	41.9	38.9	30.2	71.7	66.0	51.0	89.3	85.4	62.6	96.7	95.4	67.3	97.6	97.4	67.9
OC-B/16	149.6M	33.4	26.5	24.8	68.6	59.5	49.8	89.2	84.1	63.6	96.8	94.8	68.3	100	100	70.4
+LR-TK0	151.6M	37.4	34.4	27.4	69.0	63.0	49.9	88.8	84.2	63.4	96.8	95.1	68.4	99.0	99.0	69.8

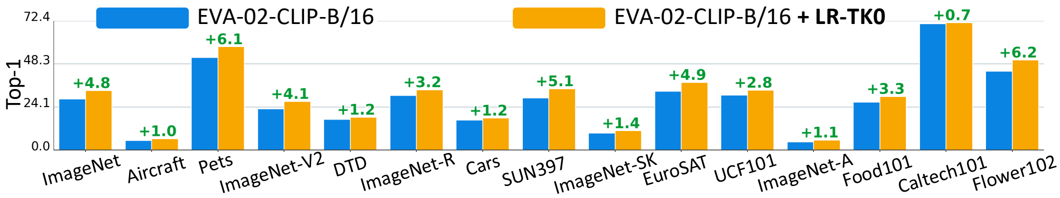

Figure 12: **Baseline vs LR-TK0**: Top-1 accuracy for EVA-B/16 on 16×16. (more in *Supplementary*)

Table 3: **Comparison with SR methods**: EVA-B/16 results, SR-specific pre-processing.

Method	16×16			32×32		
	SAR	WAR	Acc	SAR	WAR	Acc
Baseline	34.1	26.8	25.0	71.8	59.0	51.2
BSRGAN	12.4	12.2	8.8	37.3	28.7	26.9
ESRGAN	14.2	15.1	10.0	40.3	32.6	28.9
Swinir	17.9	17.6	12.7	47.7	38.3	34.3
AddSR	20.5	16.8	15.0	48.3	36.0	35.2
Inf-DiT	29.0	25.3	20.9	67.7	58.6	48.0
Our	**38.9**	**29.5**	**28.4**	**73.1**	**62.0**	**52.0**

Table 4: **Generalization of LR-TK0 with other Zero-Shot Techniques**: Visual prompt Tuning (VPT) [2022] concatenates 50 learnable tokens to spatial tokens. RobustSAM [2024] is an image segmentation model modified for classification.

LR-TK0	WAR					SAR
	16	32	64	128	224	16
Baseline	30.7	64.8	85.8	96.1	100	38.0
+VPT	35.5	64.1	84.6	94.5	97.8	42.6
+RobustSAM	32.2	61.5	82.7	92.4	93.0	37.8
+LR Tokens	35.4	66.4	85.9	95.5	98.7	42.4

6.1 Results

Table 2 shows our LR tokens consistently enhance robustness at low resolutions (16×16 & 32×32), particularly for MetaCLIP. While the low resolution is often seen as a domain shift problem (Ge et al., 2020), leading to potential declines in HR performance, our multi-scale training and HR teacher distillation minimize accuracy drops at higher resolutions (1-2% accuracy drop). Also, LR tokens have a minimal parameter gain (+3%). **Figure 12** shows Top-1 accuracy for EVA-B/16 with and without our LR-TK0, at 16×16, with max improvement on Flower-102 (6.2%). **Table 3** compares EVA-B/16 with super-resolution (SR) methods, with SR methods performing poorly in zero-shot settings for very low resolutions (fig. 9). In contrast, our approach is better suited for zero-shot scenarios. Diffusion-based SR method IDM is too computationally expensive to evaluate on large datasets like ImageNet (results in *Supplementary*). **Table 4** applies our LR-TK0 technique to visual prompt tuning which concatenates tokens (instead of adding) only before the first block. RobustSAM (segmentation models) modified for image classification (*Supplementary*).

6.2 Ablation Study

Design Choices: Table 5 shows not freezing the pre-trained weights (*i.e.* fine-tuning the last 4 blocks at 1/100 of the default learning rate) with and without LR tokens (first two rows) degrades the performance, indicating the necessity of preserving pre-trained weights. Our design choice is task agnostic *i.e.* model's classification plays no role in learning the HR-LR relationship but classifying LR images into captions (as class labels, *task-oriented*) has more or less the same performance. **Table 6** shows benefit of multi-scale training (3 buckets, faster to train).

Table 5: **Ablation:** EVA-B/16 trained with 7K captions and 50 images/caption. 'CL': use of classifier. *Not* frozen means fine-tuning end-to-end.

Frozen	LR Tk.	CL	SAR-16	WAR-16	SAR-32	WAR-32
Baseline (frozen)			38.0	30.7	74.4	64.8
			31.1	24.5	67.2	56.6
	✓		32.8	27.8	68.1	58.3
✓	✓		42.3	35.2	75.3	66.4
✓	✓	✓	42.0	34.7	75.2	65.9

Table 6: **Multi-Scale (MS) Buckets**: '+' indicates Cumulative addition. E.g. [64,128] has [16,32] and [32,64] buckets.

MS Buckets	WAR-16	WAR-32
Baseline	30.74	64.81
[16, 32]	34.01	64.77
+ [32, 64]	35.28	66.10
+ [64, 128]	35.45	**66.40**
+ [128, 224]	**35.73**	65.91

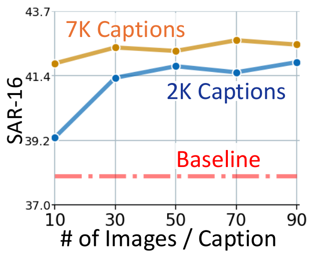

Figure 13: **#Images/Caption:** Robustness vs. Size of diffusion generated dataset.

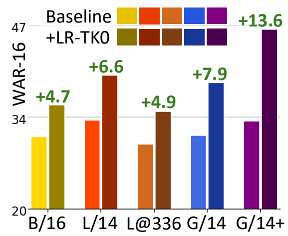

Figure 14: **LR-TK0 improves all EVA backbones**: L@336 is L/14 with 336 input

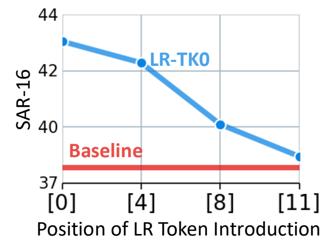

Figure 15: [i] LR tokens introduced starting from i^{th} block (& none after patchification).

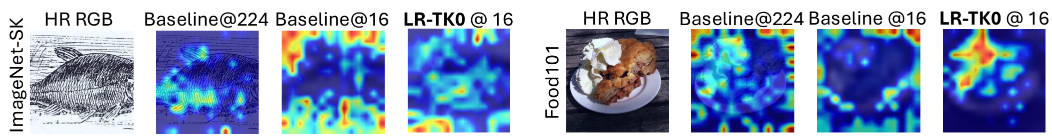

Figure 16: **LR token Grad-CAM**: Baseline (EVA-B/16) attention is scattered at 16×16 (compared to 224×224). LR-TK0 focuses on the object, likely capturing fine-grained details. @: input resolution.

 # Images/Caption: Figure 13 shows multiple images per caption & even 2000 captions consistently improve performance across 15 datasets, hinting at bridging the gap between HR-LR domains.

EVA backbones: Figure 14 shows LR tokens enhance various EVA backbones, namely, Base (B/16), Large (L/14 & L@336), and G (G/14 & G/14+). Larger backbones, B<L<G, benefit from more tokens (via more layers). Model with 336×336 input underperforms (validation, fig. 7 (left)).

Position of LR Tokens: Figure 15 shows introducing tokens in the earlier layer (starting from [i]-th block, and subsequent layers) is more helpful than later. This helps validate the observation in fig. 7 (right), *i.e.* initial layers suffer more at low resolution than deeper ones, validating the choice of fixing (introducing tokens) at initial layers than just at final features.

Grad-CAM results: On low resolutions of 16×16, vanilla model attention is dispersed and not as concentrated as 224×224 **(fig. 16)**. However, our method (w/ LR tokens) shows focus on the object which helps to learn better representations at low resolution.

7 Conclusion

Our extensive evaluation of Visual-Language Foundation Models through the LR0.FM benchmark has highlighted critical limitations in their ability to generalize under low-resolution conditions, a prevalent issue in real-world scenarios. While larger models and higher-quality pre-training datasets offer increased robustness, our findings underscore the significant impact of fine-tuning and input resolution on performance. Importantly, we observed that low-resolution inputs primarily disrupt the early layers of these models, leading to degraded performance. To address these challenges, we introduced the LR-TK0 strategy, which improves model robustness to low-resolution inputs without altering pre-trained weights, offering a practical solution for real-world applications. Additionally, our proposed Weighted Aggregated Robustness metric provides a more comprehensive evaluation of model resilience, addressing the limitations of existing metrics.

8 Acknowledgment

The authors thank Steven Dick (UCF High-Performance Computing) and Rohit Gupta (UCF CRCV) for their help in generating the synthetic data set.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Eytan Bakshy, Lili Dworkin, Brian Karrer, Konstantin Kashin, Ben Letham, Ashwin Murthy, and Shaun Singh. Ae: A domain-agnostic platform for adaptive experimentation. In *NeurIPS Systems for ML Workshop*, 2018. URL http://learningsys.org/nips18/assets/papers/87CameraReadySubmissionAE%20-%20NeurIPS%202018.pdf.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 52–68. Springer, 2016.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023a.

Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4081–4091, 2024.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=mWVoBz4W0u.

Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8215–8222, 2019.

Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 605–621. Springer, 2019.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1634–1643, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

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

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=dVaWCDMBof.

Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10021–10030, 2023.

Shiming Ge, Shengwei Zhao, Chenyu Li, Yu Zhang, and Jia Li. Efficient low-resolution face recognition via bridge distillation. *IEEE Transactions on Image Processing*, 29:6898–6908, 2020. doi: 10.1109/TIP.2020.2995049.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. *Caltech Technical Report*, 2007.

Qingpei Guo, Furong Xu, Hanxiao Zhang, Wang Ren, Ziping Ma, Lin Ju, Jian Wang, Jingdong Chen, and Ming Yang. M2-encoder: Advancing bilingual image-text understanding by large-scale efficient pretraining, 2024. URL https://arxiv.org/abs/2401.15896.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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. *CVPR*, 2021b.

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, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Syed Safwan Khalid, Muhammad Awais, Zhen-Hua Feng, Chi-Ho Chan, Ammarah Farooq, Ali Akbari, and Josef Kittler. Resolution invariant face recognition using a distillation approach. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):410–420, 2020. doi: 10.1109/TBIOM.2020.3007356.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Tin Kramberger and Božidar Potočnik. Lsun-stanford car dataset: enhancing large-scale car image datasets using deep learning for usage in gan training. *Applied Sciences*, 10(14):4913, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9287–9301. Curran Associates, Inc., 2022a.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

Pei Li, Loreto Prieto, Domingo Mery, and Patrick Flynn. Face recognition in low quality images: A survey. *arXiv preprint arXiv:1805.11519*, 2018.

Pei Li, Loreto Prieto, Domingo Mery, and Patrick J. Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019. doi: 10.1109/TIFS.2018.2890812.

Runze Li, Dahun Kim, Bir Bhanu, and Weicheng Kuo. RECLIP: Resource-efficient CLIP by training with small images. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856. URL https://openreview.net/forum?id=Ufc5cWhHko.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. In *NeurIPS*, 2023b.

Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1% zero-shot imagenet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*, 2023c.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2022.

Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou, and Delong Chen. Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, 57(10):268, 2024.

Yuanyuan Liu, Fan Tang, Dengwen Zhou, Yiping Meng, and Weiming Dong. Flower classification via convolutional neural network. In *2016 IEEE International Conference on Functional-Structural Plant Growth Modeling, Simulation, Visualization and Applications (FSPMA)*, pp. 110–116. IEEE, 2016.

Luis S. Luevano, Leonardo Chang, Heydi Méndez-Vázquez, Yoanna Martínez-Díaz, and Miguel González-Mendoza. A study on the performance of unconstrained very low resolution face recognition: Analyzing current trends and new research directions. *IEEE Access*, 9:75470–75493, 2021. doi: 10.1109/ACCESS.2021.3080712.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

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

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

Go Ohtani, Ryu Tadokoro, Ryosuke Yamada, Yuki M Asano, Iro Laina, Christian Rupprecht, Nakamasa Inoue, Rio Yokota, Hirokatsu Kataoka, and Yoshimitsu Aoki. Rethinking image super-resolution from training data perspectives. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Jyoti S. Patil, R. S. Pawase, and Yogesh H. Dandawate. Classification of low resolution astronomical images using convolutional neural networks. *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 1168–1172, 2017. URL https://api.semanticscholar.org/CorpusID:19086730.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. Probing conceptual understanding of large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1797–1807, 2024a.

Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14698–14708, 2023.

Madeline Chantry Schiappa, Shehreen Azad, Sachidanand Vs, Yunhao Ge, Ondrej Miksik, Yogesh S Rawat, and Vibhav Vineet. Robustness analysis on foundational segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1786–1796, 2024b.

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

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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, and Dimitris Metaxas. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11953–11962, October 2023.

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

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

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.

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, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2023b.

Praveen Tirupattur, Aayush J Rana, Tushar Sangam, Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Tinyaction challenge: Recognizing real-world low-resolution activities in videos. *arXiv preprint arXiv:2107.11494*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Magneto: A foundation transformer. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36077–36092. PMLR, 23–29 Jul 2023.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4582–4591, 2017.

Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=5BCFlnfE1g.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023.

Zhenlin Xu, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joe Tighe, and Davide Modolo. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *CVPR 2024 Workshop on "What is Next in Multimodal Foundation Models?"*, 2024b.

Zhuoyi Yang, Heyang Jiang, Wenyi Hong, Jiayan Teng, Wendi Zheng, Yuxiao Dong, Ming Ding, and Jie Tang. Inf-dit: Upsampling any-resolution image with memory-efficient diffusion transformer. In *European Conference on Computer Vision*, pp. 141–156. Springer, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=Ee277P3AYC.

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 (ICCV)*, pp. 11975–11986, October 2023.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pp. 4791–4800, 2021.

Wenbo Zhang, Yifan Zhang, Jianfeng Lin, Binqiang Huang, Jinlu Zhang, and Wenhao Yu. A progressive framework of vision-language knowledge distillation and alignment for multilingual scene. *arXiv preprint arXiv:2404.11249*, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

Jizhe Zhou, Chi-Man Pun, and Yu Tong. Privacy-sensitive objects pixelation for live video streaming. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 3025–3033, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413972. URL https://doi.org/10.1145/3394171.3413972.