

# The 1st LoViF Challenge on Efficient VLM for Multimodal Ad Creative Quality Scoring: Methods and Results

Jusheng Zhang<sup>†,\*</sup> Qinhan Lyu<sup>†,\*</sup> Sizhuo Ma\* Rick Cao\* Jian Wang\*  
Xin Li\* Yongsen Zheng\* Keze Wang<sup>\*,‡</sup>

Jing Yang Hong Zhang Shichao Zhang William Leach Ru He  
Min Cao Yizhen Jia Yin-Loon Khor Yi Jie Wong Peimeng Sui Yu Hao  
Weixi Lin Weijian Deng  
zhangjusheng19981128@gmail.com

## Abstract

*This paper reviews the 1st Challenge on Efficient Vision Language Models for Multimodal Creative Quality Scoring, held in conjunction with the LoViF Workshop at CVPR 2026. As VLMs become increasingly powerful, their computational efficiency remains a critical bottleneck for real-world deployment. This challenge aims to establish a benchmark for the trade-off between prediction accuracy and computational efficiency. Participants were tasked with predicting the average rating of products based on multimodal inputs (images and texts) sourced from the Amazon Reviews 2023 dataset. We evaluate the proposed solutions comprehensively based on accuracy metrics (e.g., Pearson Correlation Coefficient) and efficiency metrics (Parameters, FLOPs, and Latency). In this report, we detail the challenge setup, dataset, and evaluation metrics, and summarize the highly optimized and lightweight VLM architectures proposed by the top-performing teams.*

## 1. Introduction

The recent emergence of Generative Foundation Models and Vision Language Models [1, 2] has driven a profound paradigm shift in computer vision and multimodal understanding. By aligning visual representations with rich semantic language spaces, state of the art VLMs have demonstrated unprecedented capabilities in complex tasks [3, 4], ranging from visual question answering to zero-shot image classification. Among these applications, multimodal creative quality scoring which involves comprehensively evaluating products or creative content based on images, videos, and descriptive texts has emerged as a crucial task with immense practical value in e-commerce, advertising, and rec-

ommendation systems.

However, while modern VLMs exhibit remarkable accuracy and reasoning skills, their massive scale inherently limits their real world applicability. Leading foundation models typically rely on billions of parameters, resulting in intensive memory footprints and high inference latency [5, 6]. In practical deployment scenarios, such as processing millions of product queries on e-commerce platforms, strict constraints are placed on computational budgets. Consequently, there is an urgent need to bridge the gap between heavy, resource-intensive models and the demand for lightweight, deployable solutions.

To address this critical bottleneck, the 1st Workshop on Low-level Vision Frontiers (LoViF) introduced the Challenge on Efficient VLM for Multimodal Creative Quality Scoring, held in conjunction with CVPR 2026. Organized in collaboration with Snap Inc. and Nanyang Technological University, this challenge aims to encourage researchers and practitioners to push the boundaries of lightweight yet powerful VLMs. Specifically, participants were tasked with predicting the average rating of products using multimodal inputs sourced from the large-scale Amazon Reviews 2023 dataset [7]. The core objective is to develop efficient architecture designs such as optimized single/dual-encoder schemes, modality fusion strategies, and knowledge distillation that can maintain robust multimodal scoring capabilities without incurring prohibitive computational costs.

The challenge explicitly emphasizes the trade-off between model performance and computational efficiency. Submissions were rigorously evaluated not only on prediction accuracy (measured by the Pearson Linear Correlation Coefficient, PLCC) but also on stringent efficiency metrics, including the number of parameters, Floating Point Operations (FLOPs), and inference latency.

This challenge is held with the LoViF Workshop , con-

<sup>\*</sup> Challenge organizers. Other authors are participants in this challenge.

<https://lovif-cvpr2026-workshop.github.io/>

taining series of challenges on: real-world all-in-one image restoration [8], efficient VLM for multimodal creative quality scoring [9], weather removal in videos [10], holistic quality assessment for 4D world model [11], and human-oriented semantic image quality assessment [12].

The remainder of this report is organized as follows. Section 2 details the challenge setup, including the dataset construction and evaluation metrics. Section 3 presents the quantitative results and the final leaderboard. Section 4 provides brief descriptions of the highly optimized VLM architectures proposed by the top-performing teams. Finally, Section 5 concludes the report and discusses future directions for efficient multimodal understanding.

## 2. Challenge Setup

In this section, we provide a detailed overview of the challenge task, the dataset constructed for multimodal creative quality scoring, and the comprehensive evaluation metrics designed to assess the trade-off between model performance and efficiency.

### 2.1. Task Definition and Dataset

The primary objective of the Efficient VLM Challenge is to predict the *Average Rating* of a product given its associated multimodal information.

To establish a robust benchmark, we constructed the challenge dataset based on the **Amazon Reviews 2023 dataset**. This dataset provides a diverse and large-scale collection of multimodal product data. For each product instance, the input data fields are structured as follows:

- **Textual Attributes:** `title` (product name), `description` (detailed text), `features` (bullet-point highlights), `details` (dictionary of materials, brand, etc.), and `main_category`.
- **Visual Attributes:** `images` (multi-resolution product images) and `videos` (including titles and URLs).
- **Metadata:** `price`, `store`, `categories` (hierarchical structure), `rating_number`, `parent_asin`, and `bought_together`.

The target variable is the `average_rating` (float) shown on the product page. To ensure fair evaluation, the ground-truth ratings for the test set were strictly hidden from the participants during the competition.

### 2.2. Evaluation Protocol and Metrics

A distinguishing feature of this challenge is its strict focus on deployability. To ensure a fair comparison across diverse architectures, we enforce a **Canonical Inference Protocol**: input images are resized to a fixed resolution, text inputs are truncated to a fixed length, and only a single forward pass is allowed (no test-time augmentation or ensembling).

The final ranking is determined by the **Comprehensive**

**Efficiency Score (CES)**, which balances prediction accuracy and computational resource cost.

#### 1. Accuracy (PLCC<sup>+</sup>):

We use the Pearson Linear Correlation Coefficient (PLCC) to measure the correlation between the predicted scores and the ground truth. To avoid incentivizing negatively correlated predictions, the score is strictly clipped at 0:

$$\text{PLCC}^+ = \max(0, \text{PLCC}). \quad (1)$$

#### 2. Geometric Resource Cost (C):

To rigorously assess efficiency, we penalize imbalances between storage (Parameters) and computation (FLOPs) using a geometric mean. Let  $P$  and  $F$  denote the parameters (in Millions) and FLOPs (in GigaFLOPs) of the proposed model. The resource cost  $C$  is defined as:

$$C = \left(\frac{P}{P_{\text{tgt}}}\right)^w \left(\frac{F}{F_{\text{tgt}}}\right)^{1-w}, \quad (2)$$

where  $P_{\text{tgt}} = 1000\text{M}$  and  $F_{\text{tgt}} = 20\text{G}$  are the baseline reference targets. We set a balanced weight  $w = 0.5$ .

#### 3. Asymmetric Efficiency Factor (E):

We introduce an asymmetric incentive structure to encourage lightweight design. It applies a bounded bonus for efficient models ( $C \leq 1$ ) and a steep penalty for over-budget models ( $C > 1$ ):

$$\mathcal{E} = \begin{cases} \min(1 + \gamma, 1 + \alpha(1 - C)), & \text{if } C \leq 1 \\ \max(0, 1 - \beta(C - 1)), & \text{if } C > 1 \end{cases} \quad (3)$$

where the bonus sensitivity  $\alpha = 0.05$ , the penalty sensitivity  $\beta = 2.0$ , and the maximum bonus cap  $\gamma = 0.10$  (i.e., maximum 1.10× multiplier).

#### 4. Comprehensive Efficiency Score (CES):

The final leaderboard ranking is determined by the CES, formulated as the product of the accuracy and the efficiency factor:

$$\text{CES} = \text{PLCC}^+ \times \mathcal{E}. \quad (4)$$

## 3. Challenge Results

The Efficient VLM for Multimodal Creative Quality Scoring Challenge attracted widespread attention from the community. A total of 83 participants registered for the competition, resulting in 473 submissions during the development phase. In the final testing phase, multiple valid submissions were successfully evaluated on the hidden test set under strict canonical inference protocols. It is worth noting that submissions violating the challenge rules, such as those involving data leakage (e.g., incorporating the test set into the training phase), were strictly disqualified to ensure absolute fairness.

### 3.1. Final Leaderboard

The quantitative results of the final testing phase are summarized in Table 1. The valid submissions are ranked based on the Comprehensive Efficiency Score (CES), which serves as the primary metric balancing prediction accuracy (PLCC<sup>+</sup>) and computational resource cost ( $\mathcal{C}$ ).

### 3.2. Results Analysis

As shown in the official leaderboard (Table 1), the champion team, **olacnqoddchl**, achieved the highest CES of 0.43. They successfully delivered the most competitive prediction accuracy (PLCC<sup>+</sup> of 0.41) among all valid entries, while maintaining a highly efficient architecture. Their model utilized only 152.12M parameters and 31.14G FLOPs ( $\mathcal{C} = 0.49$ ), earning a 1.04× efficiency bonus. The runner-up, team **iamxredz**, followed closely with a CES of 0.42 and even fewer FLOPs (29.00G).

Furthermore, the results highlight diverse and extreme strategies in balancing accuracy and efficiency. For instance, team **yinloonkhor** (Rank 4) adopted an aggressive lightweight design, achieving the lowest resource cost ( $\mathcal{C} = 0.10$ ) with only 27.7M parameters and 6.8G FLOPs. This earned them the maximum possible efficiency bonus factor of 1.10, demonstrating a highly optimized architecture, albeit with a slight trade-off in raw prediction accuracy (PLCC<sup>+</sup> of 0.36).

Overall, the top valid submissions demonstrated that robust multimodal quality scoring can be effectively achieved with models heavily constrained well below 250M parameters, paving the way for highly efficient VLMs in real-world deployments.

## 4. Challenge Methods

In this section, we summarize the innovative strategies employed by the top-performing teams, focusing on their architectural optimizations and multimodal fusion techniques.

### 4.1. Team: olacnqoddchl

Team **olacnqoddchl**, led by Jing Yang, developed an efficient multimodal rating prediction system built on top of the SmolVLM2-256M-Video-Instruct backbone. Instead of introducing heavy additional fusion modules or dual encoders, their method preserves the native image-text interaction of the pretrained backbone and adapts the compact generative vision-language model into an efficient multimodal regressor.

**Overall Architecture:** The multimodal pipeline processes a single RGB product image resized to a fixed  $384 \times 384$  resolution. The textual input consists of metadata fields that are serialized into an instruction-style prompt with a maximum tokenizer length of 2560. To convert the generative VLM into a regressor, the team

explicitly removed the original language modeling head (`drop_lm_head=True`). They applied mask-aware mean pooling over the final multimodal hidden states to obtain a stable sample-level representation. This pooled representation is then passed through a lightweight two-layer MLP regression head. Finally, a sigmoid-based linear rescaling is applied to strictly bound the output to the valid 1.0 to 5.0 rating range.

**Training and Optimization:** The model was trained on the Amazon Reviews 23 dataset. To better align the optimization process with the challenge’s correlation-based evaluation criterion, the team engineered a composite PLCC-oriented loss function:  $\mathcal{L} = 0.5 \times \text{MSE} + 0.5 \times (1 - \rho(\hat{y}, y))$ , where  $\rho(\hat{y}, y)$  denotes the Pearson correlation coefficient. During training, the vision encoder was completely frozen (`freeze_vision=True`), and the model was fine-tuned using bf16 mixed precision to improve memory and compute efficiency. For inference, the team strictly adhered to an efficiency-oriented, single-model pipeline without any ensembles, achieving a test score of 0.41 while requiring only 228 million parameters and 61.44 GFLOPs per image.

### 4.2. Team: i am ok

Team **iamxredz** proposed *LoViFModel*, a highly efficient dual-tower multimodal fusion architecture built upon the pre-trained OpenAI CLIP-ViT-B/32 backbone[13]. Their solution elegantly balances inference speed, memory footprint, and prediction accuracy.+

**Overall Architecture:** To drastically reduce computational overhead while preserving pre-trained representation capabilities, the team employed a fine-grained layer freezing strategy. They froze the first 11 layers of both the visual and textual encoders, only thawing the final layer for fine-tuning. The visual encoder extracts 768-dimensional features from images, while the text encoder processes concatenated titles and descriptions into 512-dimensional features[14]. To address the dimensionality mismatch, linear projection layers are utilized to map both modalities into a unified 512-dimensional shared feature space.

**Multimodal Fusion:** Instead of relying on heavy cross-attention mechanisms, they implemented a lightweight fusion strategy. The aligned visual and textual tokens are concatenated into a sequence and fed into a compact 2-layer, 8-head Transformer Encoder[15]. This self-attention module efficiently captures bidirectional correlations between the modalities. Finally, the fused features are flattened and passed through a regression MLP head, followed by a Sigmoid function scaled to map the outputs to the 1-5 rating range.

**Training and Optimization:** A key contribution of their method is the design of a composite loss function combining Mean Squared Error (MSE) and Pearson Lin-

Table 1. Final official leaderboard of the Efficient VLM Challenge. Methods are ranked by the Comprehensive Efficiency Score (CES). Models with a resource cost  $\mathcal{C} \leq 1$  receive an efficiency bonus ( $\mathcal{E} \geq 1$ ), while those with  $\mathcal{C} > 1$  are heavily penalized. \* denotes the proposed baseline from Snap Inc., which is not included in the final ranking.

Rank	Team (Participant)	CES (Final) $\uparrow$	PLCC <sup>+</sup> $\uparrow$	Factor ( $\mathcal{E}$ )	Cost ( $\mathcal{C}$ ) $\downarrow$	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Runtime (s) $\downarrow$
1	<b>olacnqoddchl</b>	<b>0.43</b>	<b>0.41</b>	1.04	0.49	152.12	31.14	1.00
2	iamxredz	0.42	0.40	1.04	0.46	145.14	29.00	1.00
*	wleach	0.40	0.39	1.01	0.84	228.00	61.44	<b>0.01</b>
3	yinloonkhor	0.40	0.36	<b>1.10</b>	<b>0.10</b>	<b>27.70</b>	<b>6.80</b>	0.00*
4	ps3336	0.39	0.37	1.05	0.40	119.31	26.33	0.04
5	skywalker5165	0.39	0.37	1.05	0.40	119.31	26.33	0.04
6	monkeydminh49	0.34	0.33	1.03	0.55	184.43	32.64	1.00
7	atul-dev342	0.34	0.33	1.03	0.59	204.24	33.61	0.05
8	Shubhojit	0.30	0.29	1.03	0.50	124.72	39.55	1.00
9	Hope	0.23	0.23	1.01	0.81	186.09	70.58	0.01

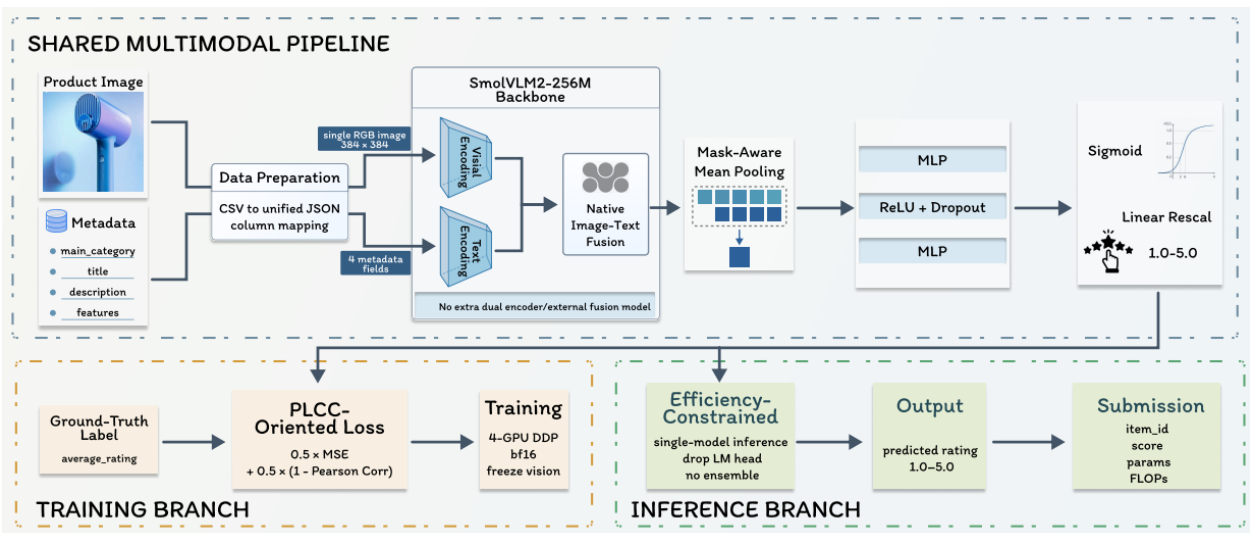


Figure 1. Overview of the efficient multimodal rating prediction pipeline proposed by team **olacnqoddchl**. The architecture leverages a shared SmolVLM2-256M backbone, preserving its native image-text fusion without introducing heavy external fusion modules. The generative language head is removed and replaced with a mask-aware mean pooling layer and a lightweight MLP regression head, bounded by a scaled sigmoid function to output a 1.0–5.0 rating. During training, a PLCC-oriented composite loss is employed, while the inference branch strictly adheres to efficiency constraints (single-model inference, dropped LM head, no ensembles).

ear Correlation Coefficient (PLCC) loss. This dual-loss setup simultaneously ensures numerical closeness to the ground truth and ranking consistency with human perception. Furthermore, their training pipeline incorporated robust optimization strategies, including a differential learning rate to stabilize backbone fine-tuning, Automatic Mixed Precision (AMP) training for memory efficiency, and a semi-supervised pseudo-labeling mechanism for dataset expansion[16].

### 4.3. Team: wleach

Team **wleach** presented an efficient adaptation of the compact SmolVLM2-256M-Video-Instruct model for direct scalar regression[17]. Their solution elegantly repurposes

a generative Vision-Language Model into a deterministic scoring architecture suitable for latency-constrained environments.

**Overall Architecture:** The team utilized a ViT-based SigLIP vision tower to extract patch-level embeddings from input images[18]. To strictly enforce computational efficiency and predictability, they resized all images to a fixed  $384 \times 384$  resolution using bilinear interpolation, completely disabling dynamic resizing and image tiling]. Both the vision encoder and its corresponding pixel-shuffle connector were kept frozen during fine-tuning. For textual and multimodal processing, they employed the SmolLM2-based decoder. Crucially, they avoided autoregressive generation at inference time by replacing the standard language mod-

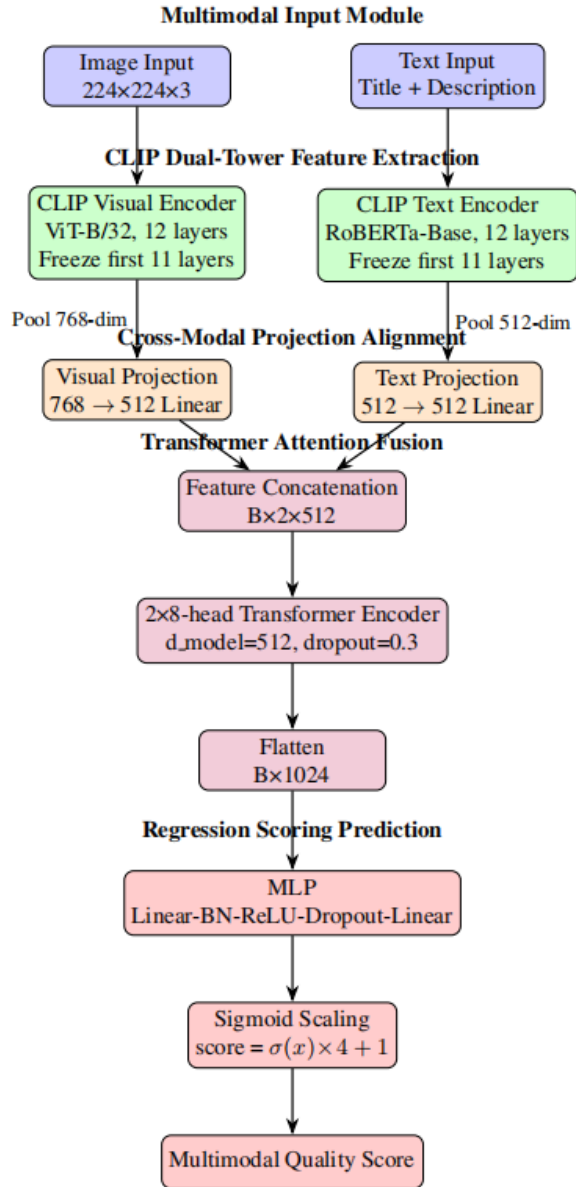


Figure 2. Overview of the *LoViFModel* architecture proposed by team **iamxredz**. The framework employs a selectively frozen CLIP dual-tower encoder (ViT-B/32 and RoBERTa-Base) to extract robust representations while minimizing computational overhead. Modality-specific projection layers align the visual and textual features into a shared 512-dimensional space. The tokens are then concatenated and fused via a lightweight 2-layer, 8-head Transformer Encoder, followed by an MLP regression head with a scaled sigmoid activation to predict the final multimodal quality score.

eling head with a lightweight two-layer MLP regression head. This head outputs a scalar value that is subsequently mapped to the 1.0 to 5.0 rating range via a scaled sigmoid transformation.

**Multimodal Fusion:** In their pipeline, the visual features are compressed via a pixel-shuffle connector and projected into the decoder embedding space before being concatenated with the tokenized text inputs. This concatenated sequence is then processed by the autoregressive decoder to produce multimodal hidden states[17]. To convert this variable-length sequence of decoder hidden states into a fixed-dimensional representation for the MLP regression head, the team implemented a mask-aware mean pooling strategy. This specific pooling mechanism explicitly excludes padding tokens via the attention mask, which significantly improves regression stability.

**Training and Optimization:** The model was fine-tuned end-to-end (excluding the frozen visual components) using a Mean Squared Error (MSE) loss. To train the model, they utilized a massive scale of approximately 16 million filtered items from the Amazon Reviews 2023 dataset. To mitigate severe category imbalance within the dataset, they applied a per-category popularity-stratified sampling strategy. Their robust optimization pipeline leveraged an 8-bit AdamW optimizer to reduce memory usage, alongside Flash Attention 2 and bfloat16 automatic mixed precision[19, 20]. This approach resulted in a highly compact model with 227.92M parameters and 61.44 GFLOPS, achieving an exceptional runtime of 0.0084 seconds per image on a single NVIDIA A100 GPU.

#### 4.4. Team: Double Y

Team **yinloonkhor** developed *EffiMiniVLM*, an exceptionally compact dual-encoder vision-language regression model that prioritizes extreme resource efficiency[21, 22]. Their architecture achieved the lowest resource cost in the challenge (0.1) by requiring merely 27.7M active parameters and 6.8 GFLOPs.

**Overall Architecture:** To achieve this minimal computational footprint, the team completely eschewed large-scale foundation models. Instead, for the visual branch, they employed a pre-trained EfficientNet-B0 to extract compact image features from  $224 \times 224$  normalized inputs. For the textual branch, product metadata was processed using a highly compressed MiniLMv2-L6-H384 transformer encoder (distilled from BERT-Large) to produce dense semantic embeddings[22].

**Multimodal Fusion:** Maintaining their strict adherence to efficiency, the team avoided any heavy cross-attention modules or ensembling strategies. The visual and textual embeddings are simply concatenated to form a joint multimodal representation. This concatenated vector is then directly passed to a lightweight Multi-Layer Perceptron (MLP) regression head to predict the final scalar product quality score.

**Training and Optimization:** Due to hardware limitations, the model was trained on a 20% subset of the provided

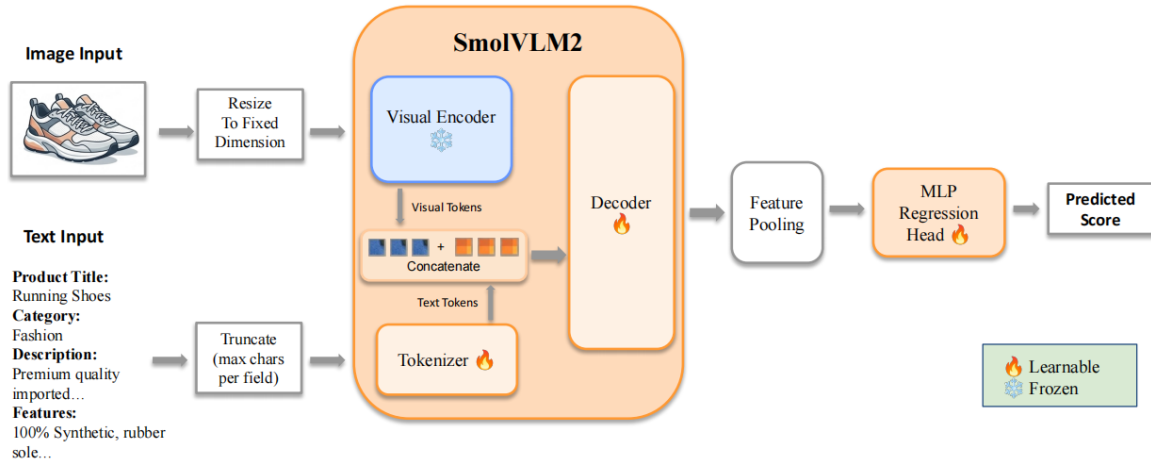


Figure 3. Overview of the architecture proposed by team **wleach**. The model leverages a fixed  $384 \times 384$  bilinear interpolation for product images and concatenated textual descriptions as inputs. The frozen SigLIP vision tower and pixel shuffle connector project visual features into the multimodal decoder space, where they are fused with text tokens by the SmoLM2 autoregressive decoder. A specialized regression pipeline replaces the language head, featuring mask-aware mean pooling to aggregate hidden states, followed by a 2-layer MLP with a scaled Sigmoid transformation to predict the final quality score.

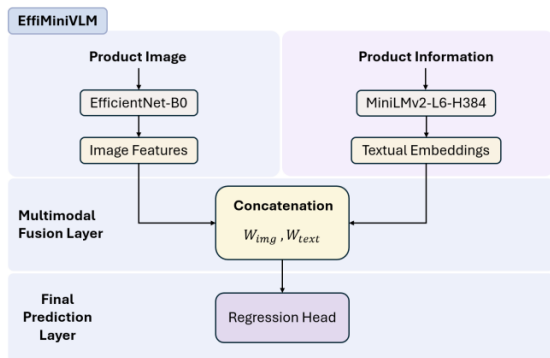


Figure 4. Overview of the *EffiMiniVLM* architecture proposed by team **yinloonkhor** (Double Y). Prioritizing extreme resource efficiency, the dual-encoder framework extracts image features using a lightweight EfficientNet-B0 and textual embeddings via a MiniLMv2-L6-H384 encoder. The unimodal representations are directly concatenated at the multimodal fusion layer and subsequently passed through a final regression head to predict the quality score.

Amazon Reviews 2023 dataset (approximately 1.6 million instances) without relying on any external data. A significant contribution of their approach is the design of a rating-count-based weighted Huber loss. This custom loss function computes per-sample weights based on the logarithm of the product’s total rating number, effectively placing greater optimization emphasis on items with more reliable, high-volume user feedback.

#### 4.5. Team: PinCO

Team **ps3336** presented a dual-encoder vision-language architecture designed to maximize efficiency through offline knowledge distillation and late fusion[23, 24].

**Overall Architecture:** To avoid the quadratic FLOPs scaling typical of unified VLMs, the team utilized independent, lightweight pre-trained encoders. The vision branch uses Swin-v2-Tiny to process  $256 \times 384$  product image collages, while the text branch relies on DeBERTa-v3-small to encode up to 256 tokens of concatenated text. The deployed student model contains approximately 179.21M parameters and requires 26.33 GFLOPs per sample.

**Multimodal Fusion:** For cross-modal reasoning, they proposed a specialized 2-layer Transformer fusion module. It first performs cross-attention where text queries attend to vision keys and values. This is followed by self-attention over a compact 5-token sequence that cleverly incorporates learned category embeddings and auxiliary numeric features. The final regression is handled by a 2-layer MLP, with the output strictly bounded to the 1.0 to 5.0 rating range using a scaled sigmoid activation.

**Training and Optimization:** A standout feature of their pipeline is offline progressive knowledge distillation. They fine-tuned a massive 2.1B-parameter Qwen3-VL-Embedding teacher model to generate soft labels and embeddings, which the compact student model learned to mimic without incurring any inference-time cost. The network was optimized using a custom multi-component loss combining MSE, Pearson correlation, and pairwise ranking, alongside inverse-frequency sample weighting to address

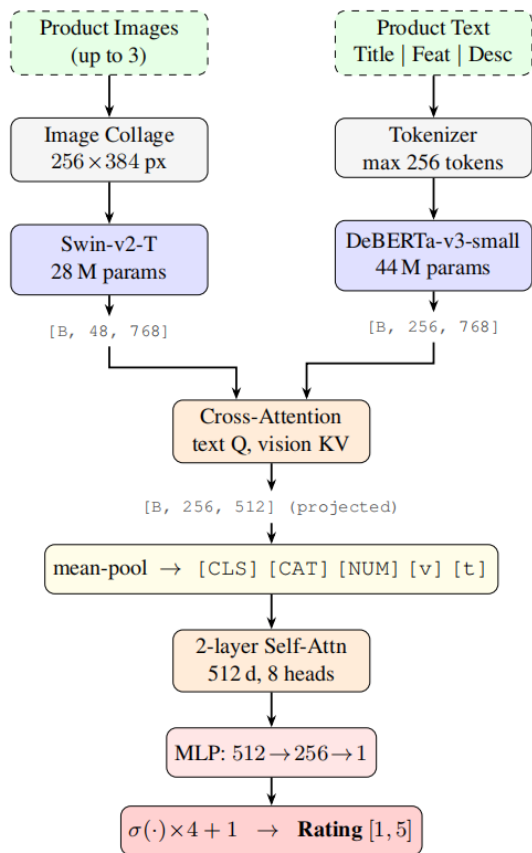


Figure 5. Overview of the dual-encoder architecture proposed by team **ps3336** (PinCO). The pipeline processes up to three product images into a  $256 \times 384$  collage encoded by Swin-v2-T, while textual metadata is encoded by DeBERTa-v3-small. Cross-modal reasoning is efficiently performed via a Cross-Attention module (text queries, vision keys/values), followed by a 2-layer Self-Attention block operating over a compact 5-token sequence. Finally, an MLP regression head predicts the quality score, strictly bounded to the  $[1, 5]$  range.

label skewness. Finally, they applied “Model Soup” by averaging the weights of three independently trained student checkpoints, securing an ensemble performance boost with zero additional runtime penalty.

#### 4.6. Team: Hope

Team **Hope** developed a scalable, single-model approach named LatentQuery Fusion, designed for hardware-agnostic efficiency[25].

**Overall Architecture:** They employed a DINOv2-base vision backbone and a lightweight DistilBERT-base-uncased text encoder[26]. To minimize the inference footprint, they disabled the video branch and utilized top- $k$  salient image selection ( $k = 2$ ) at a  $224 \times 224$  resolu-

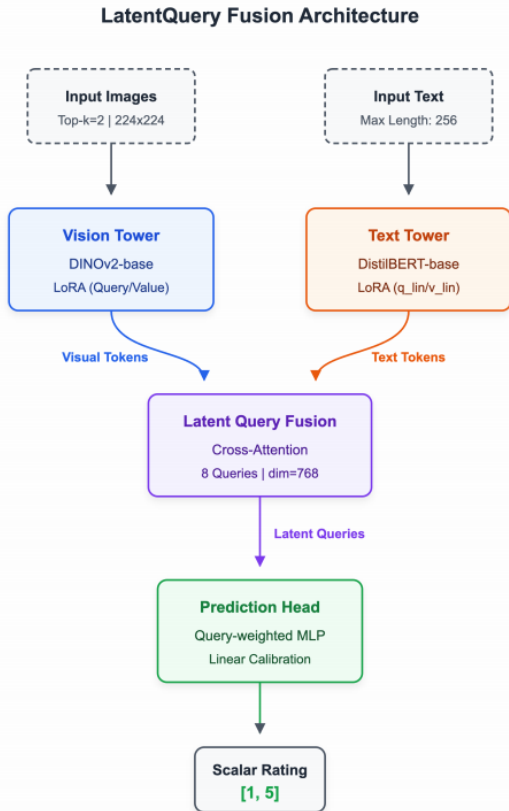


Figure 6. Overview of the LatentQuery Fusion architecture proposed by team **Hope**. The model processes up to two  $224 \times 224$  images via a DINOv2-base vision tower and text up to 256 tokens via a DistilBERT-base text tower. Both encoders are efficiently fine-tuned using Low-Rank Adaptation (LoRA) on their respective query and value projection layers. The resulting multimodal tokens are integrated using a Cross-Attention Latent Query Fusion module equipped with 8 latent queries. The fused representations are finally passed to a query-weighted MLP prediction head with linear calibration to output a scalar rating between 1 and 5.

tion. Crucially, to ensure parameter efficiency, they applied Low-Rank Adaptation (LoRA) to the attention projections of both encoders. This strategy restricted trainable parameters to merely 33.15M out of the 186.09M total parameters.

**Multimodal Fusion:** To integrate heterogeneous inputs efficiently, they introduced a LatentQuery mechanism. This block uses 8 latent queries across 3 transformer layers to decouple input token density from the regression head’s complexity. The resulting queries are passed into a query-weighted MLP followed by linear calibration to produce the final scalar rating.

**Training and Optimization:** The model was trained exclusively on the official LoViF dataset without external data. They optimized the network using AdamW and a composite objective function. The primary loss was SmoothL1 re-

gression, heavily augmented with auxiliary terms including a Pearson-oriented penalty, consistency loss, and distribution KL divergence terms. By avoiding complex ensembles, they maintained a predictable compute budget of 70.58 GFLOPs per sample when  $k = 1$ .

## 5. Discussion and Conclusion

The 1st LoViF Challenge on Efficient VLM for Multimodal Creative Quality Scoring successfully established a rigorous benchmark for deploying multimodal models under strict computational constraints. By analyzing the top-performing submissions, several distinct technical trends and valuable insights have emerged for the future of efficient multimodal understanding:

**1. Architectural Paradigms: Dual-Encoders and Late Fusion.** While massive unified Vision-Language Models (VLMs) dominate general-purpose tasks, the challenge results demonstrated that decoupled dual-encoder architectures (e.g., pairing CLIP, Swin, or EfficientNet with lightweight language models like DeBERTa or MiniLM) are exceptionally well-suited for efficiency-constrained regression tasks. By freezing early layers and utilizing lightweight late-fusion mechanisms (such as shallow Transformer encoders or latent queries), participants successfully bypassed the quadratic computational overhead typical of deep cross-attention layers.

**2. Advanced Optimization: PEFT and Knowledge Distillation.** To maximize representational capacity within a bounded parameter budget, top teams heavily relied on advanced optimization strategies. Parameter-Efficient Fine-Tuning (PEFT), particularly Low-Rank Adaptation (LoRA), allowed models to leverage powerful foundation models (like DINOv2) while training only a fraction of the parameters. Furthermore, offline knowledge distillation proved highly effective; by transferring dark knowledge from massive teacher models (e.g.,  $\approx 2$ B parameters) to compact student models ( $\approx 150$ M parameters), participants significantly boosted inference accuracy without adding any runtime latency.

**3. Objective Alignment: Custom Loss Functions.** Traditional Mean Squared Error (MSE) often falls short in quality scoring tasks where human perceptual ranking is paramount. The most successful submissions engineered composite loss functions that directly optimized the challenge metrics. By integrating Pearson Linear Correlation Coefficient (PLCC) penalties, pairwise ranking losses, and rating-count-weighted Huber losses, these models achieved superior alignment with the ground-truth score distributions.

**Conclusion.** The solutions presented in this challenge highlight that competitive multimodal quality scoring does not strictly require billion-parameter foundation models. With careful architectural design, offline distillation, and

metric-aligned optimization, robust performance can be achieved with models well under 250M parameters, and in extreme cases, under 30M parameters. These advancements pave the way for real-world deployment of highly efficient VLMs in latency-sensitive and resource-constrained industrial applications, such as edge-device recommendation systems and high-concurrency content moderation.

Future iterations of this challenge will explore stricter dynamic efficiency constraints, broader multimodal inputs (including high-framerate videos), and cross-hardware robustness to further push the boundaries of deployable low-level vision and generative AI systems.

## Acknowledgements

We would like to express our deepest gratitude to all the participants for their enthusiastic engagement, hard work, and innovative contributions to the 1st LoViF Challenge.

We extend our sincere thanks to our joint organizing partners and sponsors—Snap Inc., Sun Yat-sen University (SYSU), and Nanyang Technological University (NTU)—for their invaluable support in task design, baseline development, and providing the awards for the winning teams.

Furthermore, we acknowledge the CodaBench platform for seamlessly hosting the evaluation server and leaderboard. We are also grateful to the creators of the Amazon Reviews 2023 dataset, which provided the foundational, high-quality multimodal data that made this benchmark possible. Finally, we thank the organizers of CVPR 2026 for their comprehensive support in hosting the LoViF workshop.

## A. Teams and affiliations

### LoViF 2026 Organizers

*Title:* The 1st LoViF Challenge on Efficient VLM for Multimodal Creative Quality Scoring

*Members:*

Jusheng Zhang<sup>1</sup>,  
Qinhan Lyu<sup>2</sup> (kirinlv03@gmail.com),  
Rick Cao<sup>3</sup> (rcao@snapchat.com),  
Sizhuo Ma<sup>3</sup> (sma@snapchat.com),  
Jian Wang<sup>3</sup> (jwang4@snapchat.com),  
Xin Li<sup>4</sup>, Kaitong Cai<sup>2</sup>, Yijia Fan<sup>2</sup>, Keze Wang<sup>2</sup>,  
Yongsen Zheng<sup>1</sup>

*Affiliations:*

<sup>1</sup> Nanyang Technological University (NTU), Singapore

<sup>2</sup> Sun Yat-sen University (SYSU), China

<sup>3</sup> Snap Inc., USA

<sup>4</sup> IMCL Lab, China

### Team olacnqoddchl

*Title:* Efficient Multimodal Rating Prediction with

SmolVLM2

*Members:*

Jing Yang (Iris Young)<sup>1</sup> (yangj668@mail2.sysu.edu.cn)

*Affiliations:*

<sup>1</sup> Sun Yat-sen University, Guangzhou, China

### **Team iamxredz (iamok)**

*Title:* LoViFModel: Multimodal Fusion Architecture for Efficiency VLM Challenge

*Members:*

Hong Zhang<sup>1</sup> (zh13883408194@163.com),

Shichao Zhang<sup>1</sup> (15924942059@163.com)

*Affiliations:*

<sup>1</sup> Chongqing Jiaotong University, Chongqing, China **Team**

### **wleach**

*Title:* Bounded-Compute Multimodal Regression for Product-Rating Prediction

*Members:*

William Leach<sup>1</sup> (wleach@snapchat.com),

Ru He<sup>1</sup> (rhe@snapchat.com),

Sizhuo Ma<sup>1</sup> (sma@snapchat.com),

Yizhen Jia<sup>1</sup> (yjia@snapchat.com)

Min Cao<sup>1</sup> (mcao@snapchat.com),

Jian Wang<sup>1</sup> (jwang4@snapchat.com),

Rick Cao<sup>1</sup> (rcao@snapchat.com)

*Affiliations:*

<sup>1</sup> Snap Inc., USA

### **Team yinloonkhor (Double Y)**

*Title:* EffiMiniVLM: A Compact Dual-Encoder Regression Framework

*Members:*

Yin-Loon Khor<sup>1</sup> (yinloonkhor@gmail.com),

Yi Jie Wong<sup>2</sup> (yjwong1999@gmail.com)

*Affiliations:*

<sup>1</sup> Universiti Malaya, Kuala Lumpur, Malaysia

<sup>2</sup> Universiti Tunku Abdul Rahman, Selangor, Malaysia

### **Team ps3336 (PinCO)**

*Title:* Efficient Dual-Encoder Vision-Language Architecture for Creative Quality Prediction

*Members:*

Peimeng Sui<sup>1</sup> (psui@pinterest.com),

Yu Hao<sup>1</sup> (yhao@pinterest.com)

*Affiliations:*

<sup>1</sup> Pinterest, Inc., San Francisco, CA, USA

### **Team Hope**

*Title:* LatentQuery Fusion: Scalable Single-Model Approach

*Members:*

Weixi Lin<sup>1</sup> (weixilin@mail.nwpu.edu.cn),

Weijian Deng<sup>2</sup> (emmmvkdeng@gmail.com)

*Affiliations:*

<sup>1</sup> Northwestern Polytechnical University, China

<sup>2</sup> Beihang University, China

## **References**

- [1] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. **1**
- [2] Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities, 2026. **1**
- [3] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. **1**
- [4] 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, 2024. **1**
- [5] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. **1**
- [6] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*, 2024. **1**
- [7] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024. **1**
- [8] Xiang Chen, Hao Li, Jiangxin Dong, Jinshan Pan, Xin Li, et al. LoViF 2026 challenge on real-world all-in-one image restoration: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. **2**
- [9] Jusheng Zhang, Qinhan Lyu, Cao Sheng Wang Jian Li Xin Wang Keze Zheng Yongsun Yang Jing Ma, Sizhuo, et al. The 1st LoViF challenge on efficient vlm for multimodal creative quality scoring: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. **2**
- [10] Chenghao Qian, Xin Li, Yeying Jin, Shangquan Sun, et al. LoViF 2026 the first challenge on weather removal in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. **2**
- [11] Wei Luo, Yiting Lu, Xin Li, Haoran Li, Fengbin Guan, Chen Gao, Xin Jin, Yong Li, Zhibo Chen, et al. LoViF 2026 the first challenge on holistic quality assessment for 4d world

- model (physcore). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 2
- [12] Xin Li, Daoli Xu, Wei Luo, Guoqiang Xiang, Haoran Li, Chengyu Zhuang, Zhibo Chen, Jian Guan, and Weiping and others Li. LoViF 2026 the first challenge on human-oriented semantic image quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. 2
- [13] 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*, pages 8748–8763. PMLR, 2021. 3
- [14] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation, 2024. 3
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017. 3
- [16] Paulius Micikevicius, Sharan Narang, Jonah Alben, Justin Gregory, John Han, Nguyen Huang, Henry Kizilevich, Sergey Samet, Devin Vazquez, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 4
- [17] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 4, 5
- [18] 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*, pages 11975–11986, 2023. 4
- [19] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 5
- [20] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations (ICLR)*, 2022. 5
- [21] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2021. 5
- [22] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, 2021. 5
- [23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 6
- [24] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. 6
- [25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 7
- [26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Soyer, JVK Kinnunen, Sepehr Guzman, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.04702*, 2023. 7