UVU: Improving Multimodal Understanding via Vision-Language Unified Autoregressive Paradigm

Supplementary Material

1. Implementation Details and Analysis

1.1. Gradient Orthogonality in Image-Text Losses

We train all compared models for 1,000 steps under identical configurations and measure the average cosine angle between image-loss and text-loss gradients across the output layer, a middle layer (layer 15), and the embedding layer. This multi-layer analysis ensures fairness and robustness of the observations. The results confirm that conventional AR+VQ paradigms maintain near-orthogonal or even opposing gradients throughout training, while UVU rapidly aligns the two gradients into the same direction, fundamentally resolving the optimization conflict and enabling visual supervision to positively contribute to understanding.

1.2. Evaluation Details

We evaluate our model on the following benchmarks:

SEED-Bench [6]: A benchmark for evaluating generative comprehension in MLLMs, consisting of 19K multiple-choice questions across 12 evaluation dimensions, including both spatial and temporal understanding in images and videos.

VisuLogic [11]: A benchmark designed to evaluate visual reasoning capabilities of MLLMs, featuring 1,000 human-verified problems across six categories such as quantitative shifts, spatial relations, positional reasoning, and more.

MMStar [1]: An elite vision-indispensable multi-modal benchmark with 1,500 meticulously selected samples, designed to evaluate six core capabilities and 18 detailed axes of MLLMs.

MME [2]: A comprehensive evaluation benchmark for MLLMs, assessing both perception (coarse- and fine-grained recognition) and cognition abilities across various tasks.

CVBench-2D [9]: Part of the Cambrian Vision-Centric Benchmark (CV-Bench), focusing on 2D vision-centric Visual Question Answering (VQA) tasks, containing manually inspected examples repurposed from standard vision datasets to address limitations in existing benchmarks.

CVBench-3D [9]: Part of CV-Bench, emphasizing 3D vision-centric tasks, evaluating spatial understanding and reasoning in MLLMs using datasets like Omni3D.

VLMBlind [8]: A benchmark demonstrating that MLLMs struggle with basic visual tasks solvable by humans, such as counting, positioning, and distinguishing shapes, across seven tasks where models average 58.57% accuracy.

RefCOCO [12]: A dataset and benchmark for referring expression comprehension, providing images with annotated referring expressions to evaluate models' ability to ground language to specific objects in visual scenes.

LISA-Grounding [5]: A benchmark for reasoning segmentation tasks, comprising over one thousand image-instruction pairs that require generating segmentation masks based on complex and implicit language instructions.

ScienceQA [7]: A multimodal science question answering benchmark with approximately 21,000 multiple-choice questions covering diverse science topics, including text, images, and diagrams, to evaluate reasoning and explanation abilities.

BLINK [3]: A benchmark for assessing core visual perception abilities in MLLMs, featuring 14 tasks that humans can solve quickly but challenge current models.

HallusionBench [4]: An advanced diagnostic benchmark for evaluating entangled language hallucination and visual illusions in MLLMs, focusing on image-context reasoning with challenging samples.

It is worth noting that UVU's superior fine-grained visual perception capabilities result in its outstanding performance on visual perception-intensive tasks such as **RefCOCO, LISA-Grounding, CVBench, and BLINK.**

1.3. Large-Scale Iterative Clustering Algorithm

The codebook achieves high-fidelity reconstruction while maintaining excellent inference efficiency. We also provide a user-friendly incremental interface that seamlessly integrates with the clustering algorithm, enabling straightforward domain-specific fine-tuning and continuous incremental updates with new data.

The codebook was built from 1M images (2B 32×32×3 patches) with the following hyperparameters:

- Target codebook size $K = 200{,}000$
- Per-iteration sampling size $S = 20{,}000{,}000$
- Cache accumulation threshold $\alpha = 100{,}000{,}000$
- Long-tail threshold $\delta = 50$
- Maximum iterations $T = 50$ iterations

The complete training procedure is detailed in Algorithm 1.

1.4. Comparison with Unified Multimodal Models

Existing Unified Multimodal Models (UMMs) often incorporate visual supervision by integrating image generation

objectives into their training pipelines. However, a fundamental distinction lies in the underlying objective. While UMMs leverage visual signals primarily to enable text-to-image synthesis capabilities, our approach utilizes such supervision as a principled mechanism to bolster fine-grained visual understanding.

This distinction is not merely conceptual but is deeply rooted in the nature of the training data and its impact on model perception. As demonstrated by Wang et al. [10], the inclusion of large-scale text-to-image (T2I) datasets during the post-training stage can inadvertently degrade the model's core understanding and reasoning performance. This interference occurs because T2I data often prioritizes aesthetic qualities at the expense of semantic density, leading to a misalignment between generative proficiency and perceptual accuracy. The empirical results presented in Table 1 further corroborate this discrepancy, highlighting that the mere addition of generative objectives can be counterproductive for multimodal reasoning tasks. Furthermore, our method offers a potential pathway for high-fidelity image generation that preserves multimodal understanding.

2. More Qualitative Results

In Figure 1 to 3, we provide additional case studies.

Algorithm 1: Large-Scale Iterative Hierarchical Clustering for Pixel-Level Visual Codebook

Input: Massive patch set $\mathcal{P}_{\text{all}}$ ($\sim$2B $32\times32\times3$ patches from $\sim$1M images),
Target codebook size $K = 200{,}000$,
Per-iteration sampling size $S = 20{,}000{,}000$,
Cache accumulation threshold $\alpha = 100{,}000{,}000$,
Long-tail threshold $\delta = 50$,
Max iterations $T = 50$
Output: Pixel-level visual codebook $\mathcal{C} = \{c_k\}_{k=1}^{K}$

Extract initial batch $\mathcal{P}_0 \leftarrow 200{,}000{,}000$ patches from $\mathcal{P}_{\text{all}}$
Apply sinusoidal positional encoding to each patch vector
Randomly initialize centers $\{c_k^{(0)}\}_{k=1}^{K}$
Index all patches in $\mathcal{P}_0$ to nearest center using Faiss: $k_i = \arg\min_k \|p_i - c_k^{(0)}\|_2$
for *each cluster* k **do**
    Compute distances $d_{i,k} = \|p_i - c_k^{(0)}\|_2$ (5 decimal precision)
    Remove duplicate distances $\rightarrow$ keep only unique samples
Perform initial K-means on deduplicated data: $\mathcal{C}^{(0)} \leftarrow \text{KMeans}(\text{dedup}(\mathcal{P}_0), K)$
for $t = 1$ **to** T **do**
    Sample S new patches $\mathcal{P}_t$ from $\mathcal{P}_{\text{all}}$
    Assign each $p_i \in \mathcal{P}_t$ to nearest center in current $\mathcal{C}^{(t-1)}$ (Faiss index)
    Initialize empty cache $\mathcal{B} \leftarrow \emptyset$
    for *each cluster* k **do**
        if $|\mathcal{S}_k| < \delta$ *(long-tail)* **then**
            Cache entire $\mathcal{S}_k \rightarrow \mathcal{B}$
            mark cluster k as inactive
        else
            Compute distances to $c_k^{(t-1)}$, keep only unique distances
            Build per-cluster distance histogram and perform stratified uniform sampling (even binning from min to max distance)
            Add stratified samples to $\mathcal{B}$
    while $|\mathcal{B}| < \alpha$ **do**
        Sample additional patches, assign, dedup, stratified sample, and append to $\mathcal{B}$
    Update centers with full K-means on cached set $\mathcal{B}$: $\mathcal{C}^{(t)} \leftarrow \text{KMeans}(\mathcal{B}, K, \text{init} = \mathcal{C}^{(t-1)})$
    if *converged (center shift* $<$ *1e-4)* **then**
        break
    else
        continue
return $\mathcal{C} \leftarrow \mathcal{C}^{(t)}$

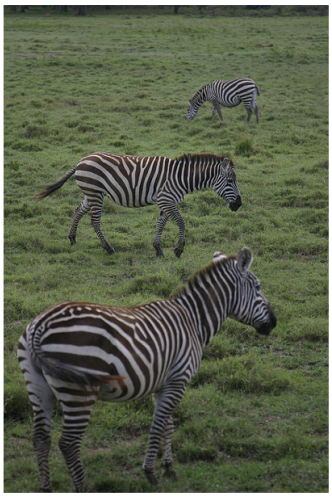
Method	# Params	LLM Backbone	SEEDB	VisuLogic	MMStar	MME	CVB2D	CVB3D	VLMBlind	RefCOCO	LISA	ScienceQA	BLINK	HallusionB
W/ Visual Supervision														
Janus-Pro	7B	DeepSeek-LLM	70.1	–	46.5	1791.7	–	–	–	–	–	83.2	38.7	39.5
Emu3	8B	LLaMA	–	–	46.4	1610.5	59.7	60.1	–	79.6	48.4	87.9	42.7	31.7
UniTok	7B	LLaMa-2	–	–	–	1448	–	–	–	–	–	–	–	–
UniCode2	7B	Qwen2.5	73.5	–	54	2052	–	–	–	–	–	–	–	–
MetaMorph	8B	LLaMA	71.8	–	–	–	–	–	–	–	–	–	–	–
show-o	1.3B	Qwen2.5	–	–	–	1097.2	–	–	–	–	–	–	–	–
Harmon	1.5B	Qwen2.5	67.1	–	–	1476	57.5	58.8	–	73.2	43.1	–	–	–
UniToken	7B	Chameleon	72.6	–	–	1922.2	–	–	–	–	–	–	–	–
TokLIP	7B	Qwen2.5	70.4	–	–	–	–	–	–	–	–	–	–	–
Chameleon	34B	LLaMa-2	48.5	–	31.8	604.5	23.4	23.1	–	51.2	29.1	58.8	21.3	18.6
UVU*	3B	Qwen2.5	73.1	24.3	52.9	2056.6	68.4	67.8	33.4	85.6	59.6	88.4	46.1	59.2
UVU(ours)	3B	Qwen2.5	**74.1**	**26.4**	**55**	**2201.9**	**73**	**76.6**	**38.8**	**91.8**	**71.7**	**91.3**	**52.8**	**66.6**

Table 1. Evaluation on multimodal understanding benchmarks. UVU* refers to the version without visual supervision.

Question:

Please describe this image in detail.

The image shows three white vases, each containing a bouquet of yellow flowers. The vases appear to be made of a textured material, possibly resembling stone or a similar substance. The flowers have a vibrant yellow color and seem to be chrysanthemums, given their layered petals and full appearance. The vases are placed on a plain, light-colored surface, which provides a neutral background that contrasts with the bright yellow of the flowers. The arrangement is simple yet elegant, with the flowers arranged in a way that allows their colors to stand out prominently.



Question:

Please describe this image in detail.

The image features a group of three zebras standing in a grassy field. The zebras are positioned at different distances from the camera, with one zebra in the foreground and two others further back in the scene.

The zebra in the foreground is closer to the left side of the image, while the other two zebras are more towards the center and right side of the frame. All three zebras appear to be grazing on the grass, enjoying their time in the open field.

Question:

Please describe this image in detail.

The image shows a group of people sitting around a couch, with a cat lying on its back in the center of the scene. The cat appears to be relaxed and comfortable, possibly enjoying the attention from the people surrounding it. There are at least three people visible in the image, with some of them holding remote controls, suggesting that they might be watching television or engaging in some other form of entertainment. The overall atmosphere seems to be casual and cozy, with everyone gathered together and sharing a pleasant moment with the cat.

Figure 1. UVU test samples on COCO dataset.

Question:
Please detect all the objects in the picture.

Figure 2. UVU test sample on COCO dataset.

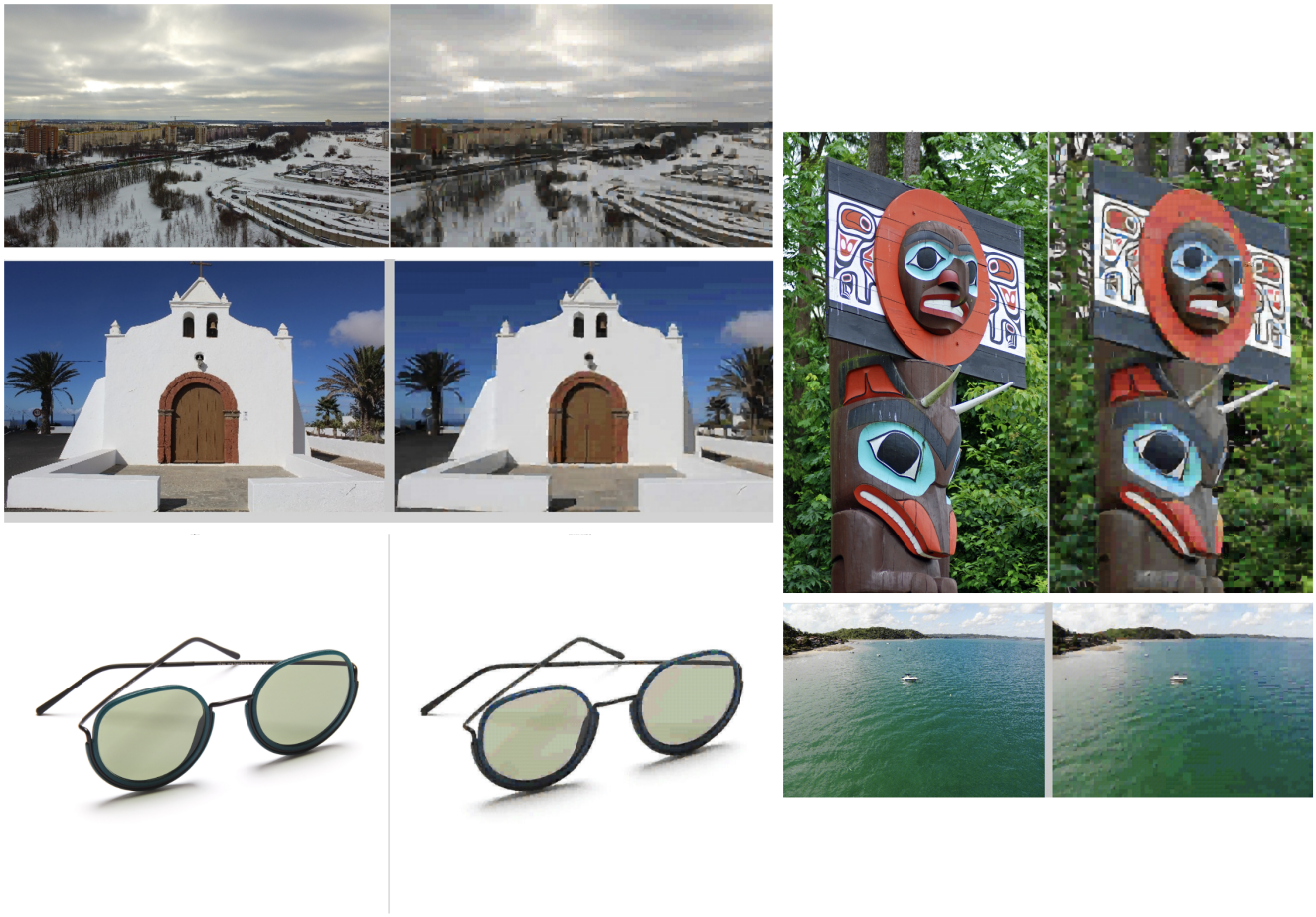


Figure 3. Comparison of the original images (**left**) and the UVU reconstructed images (**right**).

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 1

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. 1

[3] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. 1

[4] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, and Xiaoyu Liu.e.g. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. 1

[5] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model, 2024. 1

[6] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 1

[7] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. 1

[8] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. 1

[9] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, and Rob Fergus.e.g. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. 1

[10] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning, 2024. 2

[11] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025. 1

[12] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 1