Learning Vision-Language Alignment in Unified LLMs with 24 Text Tokens per Image

Nicola Irmiger[1], **Yixuan Xu**[1], **Raphael Kreft**[1], **Aram Davtyan**[2],
Manuel Kaufmann[1], **Imanol Schlag**[1],

[1]ETH Zurich, [2]University of Bern,
Correspondence: nirmiger@ethz.ch

Abstract

We explore how to adapt a pre-trained large language model to understand and generate both visual and textual information. We use an image tokenizer to compress images into discrete tokens, and train the model using the next-token prediction paradigm with the standard cross-entropy loss. A two-stage pre-training approach is applied, first training on image-only data and then on a small amount of image-text data. We evaluate how different image-text token mixing ratios during continual pre-training affect the model's ability to retain language skills while learning visual representations. The resulting model shows promising signs of flexible multimodal understanding, bridging vision and language in a single pre-trained model.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating text, as well as performing a variety of tasks (Radford et al., 2019). However, many real-world problems require reasoning over multiple modalities, such as images, videos, or audio. Vision-language foundation models have also been explored for robotic control and other embodied applications (Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2024). Training a multimodal LLM (MLLM) from scratch is extremely expensive. Extending a pre-trained LLM to other modalities while preserving language capabilities offers a more efficient path toward flexible multimodal models.

A key challenge in turning an LLM into a MLLM is preserving language capabilities while learning new modalities such as vision. Extending a text-only model can degrade text performance, a phenomenon known as catastrophic forgetting (Zhai et al., 2023b). Our work addresses this by developing a training strategy that integrates vision into a text-pre-trained model while limiting, but not entirely eliminating, degradation in language performance.

In this paper, we present a method to extend a text-pre-trained model to multimodal vision-language tasks, which can also generalize to other modalities like audio. We extend the model's embeddings and output head to accept tokens from the new modality, following Wang et al., 2026. We then study training setups varying the balance of image and text tokens in each batch to mitigate catastrophic forgetting and evaluate the model on language and vision tasks. Unlike many vision-language models that process visual features separately, we adopt a unified generative treatment of images and text, enabling learning from unpaired image data with a single training objective across modalities.

Contributions. Our main contributions are:

- **Leveraging unpaired image data.** Our approach enables large-scale pre-training on unpaired image data, reducing dependence on costly and noisy paired datasets while maintaining a unified generative framework.
- **Efficient multimodal alignment.** Effective image-text alignment can be achieved with an average of only **24 text tokens per image** in the second training stage, substantially reducing supervision and compute.
- **Analyzing language preservation.** We systematically investigate how different image-text mixing ratios affect language performance during continual pre-training, providing insights into mitigating catastrophic forgetting.

2 Related Work

Building on the success of pre-trained LLMs, recent research has developed multimodal foundation

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 275–287
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

models that integrate visual information with language. There are several approaches for enabling a model to learn a new modality, which can generally be grouped into two broad families.

Feature-based Approaches. The first family leverages the continuous features produced by a pre-trained vision encoder, such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023a). These features are integrated into the model through cross-attention layers or similar methods, allowing the model to process visual information alongside text (Deitke et al., 2025; Dai et al., 2024; Chen et al., 2024; Bai et al., 2023; Alayrac et al., 2022; Chen et al., 2023). While effective, a key limitation of this approach is that the new modality is not handled in the same generative framework as text, which prevents training on unpaired visual data and constrains flexibility across modalities. Recent work has shown that such architectures are prone to modality collapse and text dominance, where models rely heavily on textual information and largely ignore visual inputs (Wu et al., 2025a; Sim et al., 2025; Frank et al., 2021). This imbalance has been systematically measured across multiple modalities, revealing fundamental limitations in cross-modal representation learning. Additionally, it has been shown that contrastive image encoders have a tendency to overlook important visual details (Tong et al., 2024).

Token-based Generative Approaches. The second family uses discrete image tokens and a generative objective for the new modality, typically based on next-token prediction. Images are tokenized into sequences of discrete units, which are then fed into an autoregressive model alongside text tokens. This approach enables the model to directly generate visual outputs, maintaining a unified generative framework across modalities (Wang et al., 2025; Ma et al., 2025; Wang et al., 2026; Cui et al., 2025; Wu et al., 2025b; Qu et al., 2024; Team, 2025; Xie et al., 2025b,a; Jin et al., 2024). While prior works in this family primarily rely on paired image-text data, the potential to leverage abundant and inexpensive image-only data for learning internal visual representations remains underexplored.

We follow this paradigm, but explicitly emphasize a unified treatment of vision and language. By tokenizing images and training the model with a next-token prediction objective, visual and textual modalities are handled in the same generative framework, helping to reduce the tendency toward text dominance and under-utilization of visual inputs seen in some feature-based models. This design allows pre-training on both abundant unpaired images and paired image-text data while preserving language capabilities. Our experiments demonstrate that effective visual representations can be learned from unpaired images, reducing reliance on costly paired data. Moreover, the next-token prediction objective naturally enables scaling with model size and data (Kaplan et al., 2020), offering a path to further improvements.

3 Model Design

In this section, we describe our approach for extending the model to handle tokens from new modalities, and we explain our selection process for the image tokenizer, including the final choice used in our experiments.

3.1 Multimodal Large Language Model

Our approach for extending a pre-trained LLM with new tokens follows prior works such as Emu3 (Wang et al., 2026), TokenFlow (Qu et al., 2024), and Unitok (Ma et al., 2025).

Specifically, we expand the model's embedding matrix to accommodate the discrete image tokens generated by the image tokenizer, as well as special structure tokens such as `<begin-of-image>` and `<end-of-image>` markers. This approach preserves all learned embeddings for text tokens while only initializing new embeddings for image tokens.

To allow the model to predict the new tokens, we similarly expand the output layer, preserving the pre-trained weights for existing tokens and randomly initializing only the new dimensions corresponding to image tokens.

These modifications preserve the model's language performance entirely, while requiring minimal changes to the original architecture. They fully leverage the pre-trained weights, making the extension both straightforward and resource-efficient.

In our experiments we use Llama3.2-3B (Dubey et al., 2024) as the backbone model.

3.2 Image Tokenizer

Selecting an appropriate image tokenizer is not straightforward, as many options exist with varying characteristics and trade-offs (Jia et al., 2025). Our goal was to choose a tokenizer that compresses images efficiently while preserving visual information and supports images of arbitrary shapes, ensuring flexibility across different training datasets.

Image tokenizers generally fall into two categories: fixed-shape and flexible (arbitrary-shape) tokenizers. Fixed-shape tokenizers are trained for a specific input resolution and cannot inherently handle arbitrary image sizes. Flexible tokenizers inherently produce token sequences proportional to the input dimensions and can accept variable image sizes. In the following subsections, we describe each category, the tokenizers we evaluated, and the trade-offs observed.

To systematically compare tokenizers, we evaluated each candidate on its ability to accurately reconstruct images from datasets relevant to our work, including natural images, medical images, OCR data, and handwritten notes. Reconstruction quality was quantified using LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR, and token efficiency was measured as the average number of tokens per image. Tokenizers without decoders, such as some CLIP-based discrete encoders, were excluded, which we consider acceptable given their known limitations in capturing detailed visual information (Tong et al., 2024). The results are summarized in Table 1.

3.2.1 Fixed-shape Tokenizers

Fixed-shape tokenizers do not inherently support arbitrary image sizes, but we adopt resizing or tiling to handle this limitation. To leverage their efficiency while accommodating arbitrary images, we adopted a tiling strategy: each image is divided into tiles matching the expected input size of the tokenizer, with padding applied when necessary. Tiles are tokenized and reconstructed independently, and the final image is obtained by stitching the tiles together. While effective, this approach may introduce minor artifacts at tile boundaries and generates additional padding tokens, reducing overall tokenization efficiency.

Several fixed-shape tokenizers were trained for vision-language model applications and produce tokens aligned with language, which can be advantageous for multimodal vision-language tasks. Examples include TokenFlow (Qu et al., 2024) and VILA-U (Wu et al., 2025b). Selftok (Wang et al., 2025), based on reverse diffusion, produces autoregressive tokens; the larger variant achieves strong reconstruction at a high token cost, while the smaller variant trades reconstruction quality for fewer tokens. FlowMo (Sargent et al., 2025) exhibits a similar trade-off. Other fixed-shape tokenizers we considered include LlamaGen (Sun et al., 2024), DetailFlow (Liu et al., 2025), FQGAN (Bai et al., 2024), IBQ (Shi et al., 2025) (Index Backpropagation Quantization), Open-MAGVIT2 (Luo et al., 2025) (which scales the codebook to massive sizes), and VQGAN (Esser et al., 2021). Each shows trade-offs between reconstruction quality, token efficiency, and suitability for a fully discrete generative pipeline.

3.2.2 Flexible Tokenizers

Flexible tokenizers generate token sequences proportional to the input dimensions, allowing them to accept arbitrary image sizes without tiling. In our experiments, the tokenizers of the Emu model family (Emu3 (Wang et al., 2026) and Emu3.5 (Cui et al., 2025)) showed a favorable trade-off between reconstruction quality and token efficiency in our benchmarks. Unitok (Ma et al., 2025) is also flexible in input size and achieves a strong balance across our evaluation metrics. Unitok was additionally trained with alignment to CLIP/language features, but its use of multiple codebooks means it does not produce strictly single discrete tokens. Cosmos (NVIDIA et al., 2025) supports arbitrary input sizes and is available with different compression settings, but in our reconstruction evaluation it did not perform as well as some other candidates.

3.2.3 Selected Tokenizer

Based on our evaluation, we selected Emu3 as the image tokenizer for our work. Emu3 handles arbitrary image shapes natively, simplifying preprocessing and reducing artifacts caused by tiling. Additionally, it performs very well across reconstruction and token efficiency metrics, making it the most suitable option for our unified generative vision-language pipeline.

Image Token Sequence Structure. For image token structure, we follow the approach introduced in Emu3 (Wang et al., 2026), adopting their formatting scheme exactly. For each image, the Emu3 tokenizer produces a sequence of discrete tokens that are integrated into the model's input stream. Each sequence begins with a `<BOI>` token and ends with a `<EOI>` token to mark image boundaries. We prepend a small number of metadata tokens that encode the image's original width and height. Since Emu3 supports arbitrary input shapes, `<EOL>` tokens are inserted to mark line breaks, allowing the model to process images in a raster-like order. Additionally, an `<EOF>` (end of frame) token is included to reserve compatibility with video tokenization. This structured tokenization scheme allows the autoregressive model to handle visual data using the same next-token prediction objective as text, while retaining spatial structure essential for accurate reconstruction and generation.

4 Continual Pre-training Method

Our continual pre-training procedure follows a two-stage approach. In the first stage, the model is trained on large-scale unpaired images, with a small proportion of language data mixed in. To avoid the model drifting away from linguistic ability, a controlled proportion of sequences are drawn from text-only corpora and mixed into every batch. This strategy is similar to methods used for teaching a model a new language, such as Japanese (Fujii et al., 2024), where continued exposure to previously learned languages prevents forgetting. This encourages the model to develop robust internal image representations while maintaining basic language modeling capabilities. In the second stage, we use image-text pairs to explicitly align visual and textual representations. During this stage, we continue to mix in language-only sequences to prevent catastrophic forgetting of linguistic knowledge and to sustain balanced multimodal capabilities.

Compute. Each continual pre-training run required approximately 1,500 GPU hours on GH200 GPUs. The exact duration varied depending on the image-to-text token ratio per batch, as a lower proportion of image tokens results in more total tokens being consumed.

4.1 Learning Internal Image Representations

The first stage of continual pre-training is designed to expose the model to a wide distribution of visual data without relying on paired captions. Unlike most early-fusion discrete VLMs, which are trained primarily on image-text pairs, our model is initially trained exclusively on unpaired images. Captions are comparatively noisy, expensive, and difficult to collect at scale. Instead, we leverage unpaired images from diverse domains and tokenize them using the selected image tokenizer. This image-only phase allows the model to strengthen its internal representation of images while retaining useful textual priors, providing a robust foundation for subsequent multimodal alignment.

4.2 Aligning Image and Text Representations

The second stage introduces paired image-text data. Each image is tokenized into a discrete sequence and paired with its corresponding caption. This setup allows the model to align visual features with textual semantics, bridging the gap between the two modalities. Despite the overwhelming number of image tokens (on average, only 24 text tokens per image), we do not mask or down-weight the image tokens, unlike prior work. To stabilize training and preserve strong language skills, we continue to interleave independent text-only sequences throughout this stage. This ensures that improvements in multimodal alignment do not come at the expense of the model's generative language modeling performance.

4.3 Data

The datasets for each stage were selected to match the corresponding training objectives. In stage one, we rely on large-scale collections of unpaired images drawn from diverse sources (laion, 2023; OleehyO, 2024; Shao et al., 2019; Li et al., 2024; Russakovsky et al., 2015), with most datasets obtained through FineVision[1]. After tokenization, 13.2M images resulted in approximately 54B image tokens. Stage two focuses on explicit multimodal alignment and therefore uses an image-caption dataset, specifically Conceptual 12M (Changpinyo et al., 2021), that provides semantic links between modalities, yielding around 20B paired tokens after tokenization of 3.8M images. To remain within the model's context length without requiring long-context extensions, we filtered the images to be between 256×256 and 720×720 pixels, ensuring that the resulting sequences fit within the maximum length. Finally, the text-only data incorporated in both stages is drawn from FineWeb (Penedo et al., 2024). An overview of the complete data distribution is provided in Figure 1.

5 Results and Discussion

To evaluate the effect of our continual pre-training method, we conducted experiments varying the proportion of image and text tokens in each training batch. This allowed us to assess how mixing in image data impacts the model's language modeling capabilities.

All experiments were trained using standard settings, including the Adam optimizer (Kingma and

[1] https://huggingface.co/spaces/HuggingFaceM4/FineVision

Tokenizer	PSNR ↑	SSIM ↑	LPIPS ↓	Avg. #Tokens ↓
Emu3VisionTokenizer	**29.60**	0.9444	**0.0224**	10947.4
selftok_large_tiled_256	<u>28.19</u>	**0.9568**	<u>0.0230</u>	21299.2
unitok_256	26.83	0.9477	0.0263	5324.8
flowmo_hi_tiled_256	27.71	<u>0.9527</u>	0.0265	21299.2
LlamaGen_tiled_256	26.87	0.9288	0.0303	21299.2
selftok_small_tiled_256	25.60	0.9363	0.0373	10649.6
tokenflow_tiled_384	24.23	0.9162	0.0390	7435.8
flowmo_lo_tiled_256	25.33	0.9201	0.0437	5324.8
Emu3_5_IBQ	23.86	0.9140	0.0460	**2726.2**
IBQ_tiled_256	23.05	0.9120	0.0467	5324.8
detailflow_tiled_256	23.69	0.9074	0.0476	10649.6
fqgan_triple_tiled_256	24.05	0.9175	0.0490	15974.4
tokenflow_tiled_224	22.77	0.8963	0.0539	5448.8
Cosmos-0.1-Tokenizer-DI8x8	21.08	0.8665	0.0690	17750.8
vqgan_openimage_cb16384	18.86	0.8022	0.0738	17700.8
vqgan_openimage_cb256	18.43	0.8149	0.0787	17700.8
OpenMAGViT2_256	19.67	0.8303	0.0791	17814.5
vila-u_tiled_256	21.22	0.8721	0.0813	21299.2
Cosmos-0.1-Tokenizer-DI16x16	16.79	0.7424	0.1368	<u>4414.6</u>

Table 1: A quantitative evaluation of different tokenizers in terms of reconstruction quality and token efficiency is provided. The evaluation metrics include PSNR, SSIM, LPIPS, and the average number of tokens per image. The best and second-best values in each column are **bolded** and <u>underlined</u>, respectively. "CB" refers to the codebook size, and the integers represent the resolution at which the tokenizer was trained. For the Cosmos tokenizer, the compression ratios are specified.

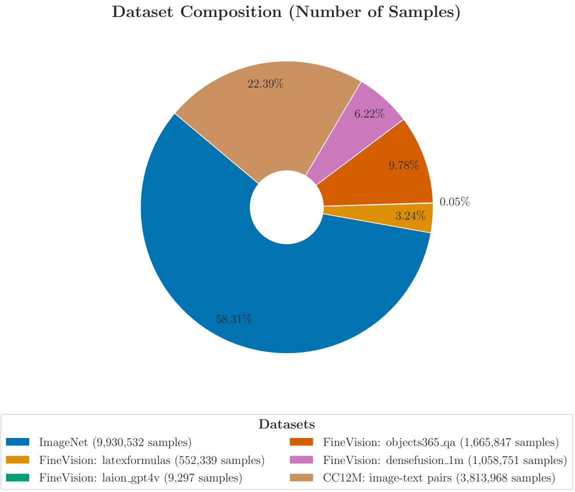

Figure 1: The distribution of training data used in the experiment. For training we only used the paired image-text data of CC12M, for the other datasets only the images were used.

Ba, 2017), without additional hyperparameter tuning. Quantitative evaluation focused exclusively on the model's performance on language modeling benchmarks to ensure that language capabilities were preserved throughout pre-training. Performance on these benchmarks was measured using the `lm-eval` framework (Gao et al., 2024).

We also provide a qualitative evaluation of the model's behavior on image-based tasks, such as completing an image or generating a caption for a given image, to illustrate its ability to leverage the newly introduced visual modalities in a generative setting.

5.1 Maintaining Language Capabilities

Figure 2 compares the performance of different models on several language pre-training benchmarks (Bisk et al., 2020; Sakaguchi et al., 2019; Zellers et al., 2019; Clark et al., 2018). As a reference point, we include the original backbone, Llama3.2-3B, shown in blue. All other models were continually pre-trained on approximately 24B image tokens, with varying proportions of text tokens mixed into each batch. As expected, higher text-to-image ratios help preserve language performance. However, increasing the share of text tokens also raises training costs, making it necessary to strike a balance. To this end, we focused on configurations with 80% and 90% image tokens per batch, which allow the model to process substantially more image data while still retaining language ability.

The evolution of performance for these two configurations is shown in Figure 3. In both cases, language performance drops sharply at the start of continual pre-training but then stabilizes. During the final two checkpoints of each stage, which coincide with the learning rate cooldown, we observe a modest recovery in language benchmark scores.

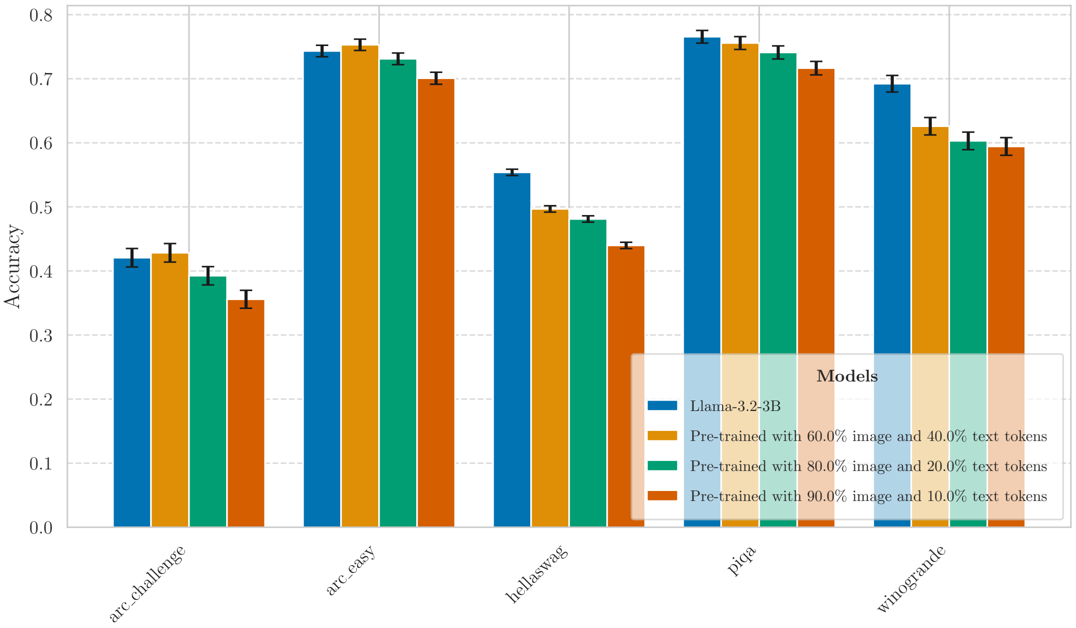

Figure 2: Comparison of benchmark performance after training with varying ratios of image and text tokens. This bar plot shows the accuracy of different models, each pre-trained with different proportions of image and text tokens, with a total of 24 billion image tokens consumed for each model. One model, the base model, is included for reference; its performance is shown alongside models pre-trained on different image-to-text token ratios. The plot compares the performance on several benchmarks, highlighting the effect of the image-token ratio on model accuracy. The error bars represent the standard deviation across multiple evaluations.

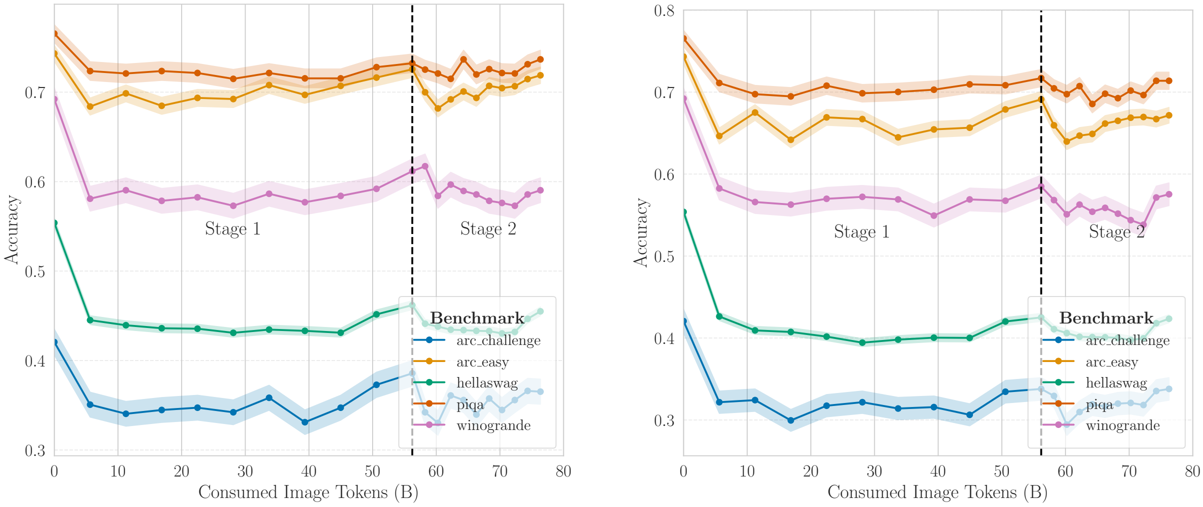

Figure 3: Training progress on benchmarks for different image/text token ratios, showing accuracy over consumed image tokens (in billions). Left: 80% image / 20% text, Right: 90% image / 10% text tokens per batch.

5.2 Qualitative Image Understanding

After verifying that language performance remains within an acceptable margin relative to the backbone model, we qualitatively assess the model's image understanding capabilities through two experiments.

In the first experiment, we provide the model with tokens corresponding to a partial image and evaluate its ability to complete the missing regions. Representative results are shown in Figure 4 and more examples in Figure 6. Several noteworthy patterns emerge. First, the model respects and reproduces the original image metadata, such as spatial dimensions, demonstrating that it has learned to generate images with arbitrary shapes, enabled by the flexible tokenizer used during training. Second, some completions indicate a non-trivial understanding of the image content. In the case of the line chart, the model not only extends the plotted line but also adds a y-axis label ("55"), indicating structural awareness. Similarly, the model successfully captures and completes the general shape of a logo, reconstructs architectural structures and surrounding trees in a building scene, and completes a checkerboard pattern without difficulty.

In the second experiment, we condition the model on the complete image followed by the prompt "the image shows" to evaluate its ability to produce descriptive captions. Example generations are shown in Figure 5 and more examples in Figure 7. The model consistently identifies visual attributes such as color and demonstrates a basic counting ability, with occasional errors. It produces a notably coherent description for the building image and identifies spatial context in the dog image, specifying that the dog is "sitting on a lawn".

Overall, these qualitative results indicate that the model develops a surprisingly solid grasp of visual structure and semantics despite being trained with a limited amount of image data and compute.

6 Conclusion

In this work, we explore a simple yet effective strategy for continual pre-training a multimodal language model by interleaving image and text data without modifying the model architecture. While discretizing images into token sequences allows images to be processed with the same generative objective as text, the central challenge we address is preserving language performance during image-heavy pre-training while also effectively leveraging large-scale image-only data. To this end, we adopt a two-stage pre-training procedure: first, the model is exposed to large amounts of unpaired image data interleaved with text to prevent catastrophic forgetting; then, paired image-text data is introduced to explicitly align the two modalities.

Through systematic tokenizer evaluation, we identified Emu3 as a highly suitable choice for our pipeline, balancing reconstruction quality and token efficiency while handling variable input sizes natively. Using this tokenizer, we continually pretrained a Llama3.2-3B backbone on roughly 74B image tokens with different image-to-text ratios per batch. Our results on standard language pretraining benchmarks show that, with careful text mixing, the model retains most of its linguistic capabilities even at high image-to-text ratios, highlighting that data composition plays a critical role in successful multimodal extension.

Qualitative experiments further indicate that meaningful visual behavior emerges, including image completion and basic caption generation, despite the relatively modest data and compute budgets used in this study.

Building on these results, we plan to scale both the training data and model size, leveraging established scaling laws to further enhance multimodal capabilities. We aim to extend the recent Apertus (Hernández-Cano et al., 2025) model with visual capabilities, leveraging our model-agnostic approach to integrate vision into existing language backbones. Following image-pre-training, we will perform visual instruction tuning (Liu et al., 2023) to improve downstream task performance and evaluate the model on vision-language benchmarks.

Our results demonstrate that large-scale image-only data can be leveraged during continual pretraining, with initial image-text alignment emerging using an average of only 24 text tokens per image. This highlights that strong multimodal abilities can arise from a lightweight pre-training pipeline, reducing reliance on costly paired datasets and providing a foundation for scaling toward more powerful vision-language models.

7 Acknowledgements

This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID infra01 on Alps. Aram Davtyan has been supported by SNSF Project 10001278.

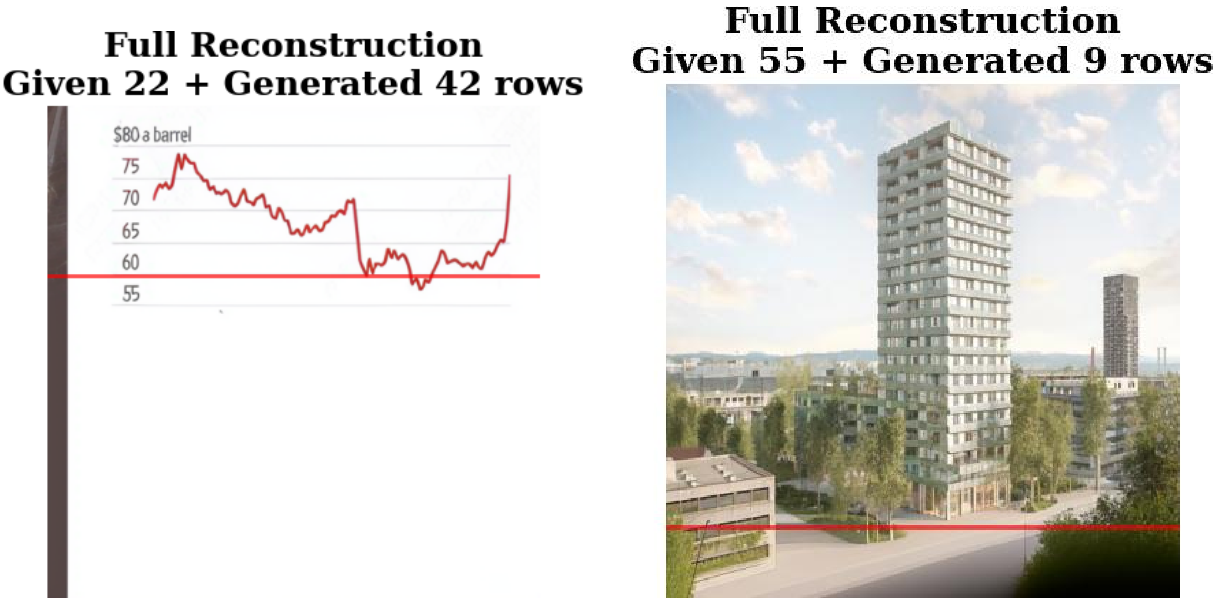

Figure 4: Image generation evaluation. The model is conditioned on the subset of image tokens above the red line and autoregressively generates the remaining tokens to reconstruct the full image.

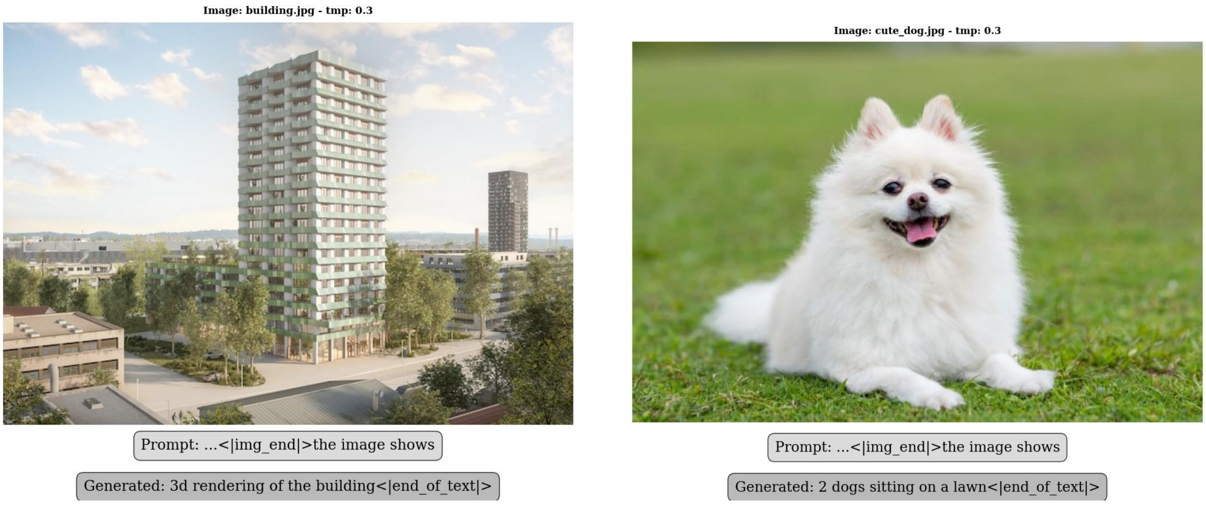
Image: building.jpg - tmp: 0.3
Image: cute_dog.jpg - tmp: 0.3

Prompt: ...<|img_end|>the image shows

Generated: 3d rendering of the building<|end_of_text|>

Prompt: ...<|img_end|>the image shows

Generated: 2 dogs sitting on a lawn<|end_of_text|>

Figure 5: Caption generation evaluation. The model is conditioned on the full set of image tokens and the sentence prefix "the image shows". It then autoregressively generates the remaining text tokens as a caption.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint*. ArXiv:2308.12966 [cs].

Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. 2024. Factorized Visual Tokenization and Generation. *arXiv preprint*. ArXiv:2411.16681 [cs].

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. `https://www.physicalintelligence.company/download/pi0.pdf`.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. 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 Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *arXiv preprint*. ArXiv:2209.06794 [cs].

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. 2025. Emu3.5: Native multimodal models are world learners. *Preprint*, arXiv:2510.26583.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NVLM: Open Frontier-Class Multimodal LLMs. *arXiv preprint*. ArXiv:2409.11402 [cs].

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Ines Altemir Marinas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonçça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. *arXiv preprint*. ArXiv:2509.14233 [cs].

Jian Jia, Jingtong Gao, Ben Xue, Junhao Wang, Qingpeng Cai, Quan Chen, Xiangyu Zhao, Peng Jiang, and Kun Gai. 2025. From Principles to Applications: A Comprehensive Survey of Discrete Tokenizers in Generation, Comprehension, Recommendation, and Information Retrieval. *arXiv preprint*. ArXiv:2502.12448 [cs].

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin CHEN, Chengru Song, dai meng, Di ZHANG, Wenwu Ou, Kun Gai, and Yadong MU. 2024. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*. ArXiv:2001.08361 [cs].

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2025. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv preprint*. ArXiv:1412.6980 [cs].

laion. 2023. gpt4v-dataset. https://huggingface.co/datasets/laion/gpt4v-dataset.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. 2024. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. In *Advances in Neural Information Processing Systems*, volume 37, pages 18535–18556. Curran Associates, Inc.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yiheng Liu, Liao Qu, Huichao Zhang, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Xian Li, Shuai Wang, Daniel K. Du, Shu Cheng, Zehuan Yuan, and Xinglong Wu. 2025. DetailFlow: 1D Coarse-to-Fine Autoregressive Image Generation via Next-Detail Prediction. *arXiv preprint*. ArXiv:2505.21473 [cs].

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. 2025. Open-MAGVIT2: An Open-Source Project Toward Democratizing Auto-regressive Visual Generation. *arXiv preprint*. ArXiv:2409.04410 [cs].

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. 2025. UniTok: A Unified Tokenizer for Visual Generation and Understanding. *arXiv preprint*. ArXiv:2502.20321 [cs].

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. 2025. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint*. ArXiv:2501.03575 [cs].

OleehyO. 2024. latex-formulas. https://huggingface.co/datasets/OleehyO/latex-formulas.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. 2024. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *CoRR*, abs/2412.03069.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Kyle Sargent, Kyle Hsu, Justin Johnson, Li Fei-Fei, and Jiajun Wu. 2025. Flow to the Mode: Mode-Seeking Diffusion Autoencoders for State-of-the-Art Image Tokenization. *arXiv preprint*. ArXiv:2503.11056 [cs].

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438. ISSN: 2380-7504.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. 2025. Scalable Image Tokenization with Index Backpropagation Quantization. *arXiv preprint*. ArXiv:2412.02692 [cs].

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs Actually See and Read? A Survey on Modality Collapse in Vision-Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint*. ArXiv:2406.06525 [cs].

Chameleon Team. 2025. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint*. ArXiv:2405.09818 [cs].

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, Mingze Zhou, Wang Lin, Kaihang Pan, Saining Zhang, Liyu Jia, Wentao Hu, Wei Zhao, and Hanwang Zhang. 2025. Selftok: Discrete Visual Tokens of Autoregression, by Diffusion, and for Reasoning. *arXiv preprint*. ArXiv:2505.07538 [cs].

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. 2026. Multimodal learning with next-token prediction for large multimodal models. *Nature*.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025a. When Language Overrules: Revealing Text Dominance in Multimodal Large Language Models. *arXiv preprint*. ArXiv:2508.10552 [cs].

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2025b. VILA-u: a unified foundation model integrating visual understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025a. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. 2025b. MUSE-VL: Modeling Unified VLM through Semantic Discrete Encoding. *arXiv preprint*. ArXiv:2411.17762 [cs].

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023a. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023b. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning (Proceedings Track)*.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning*, pages 2165–2183. PMLR. ISSN: 2640-3498.

8 Appendix

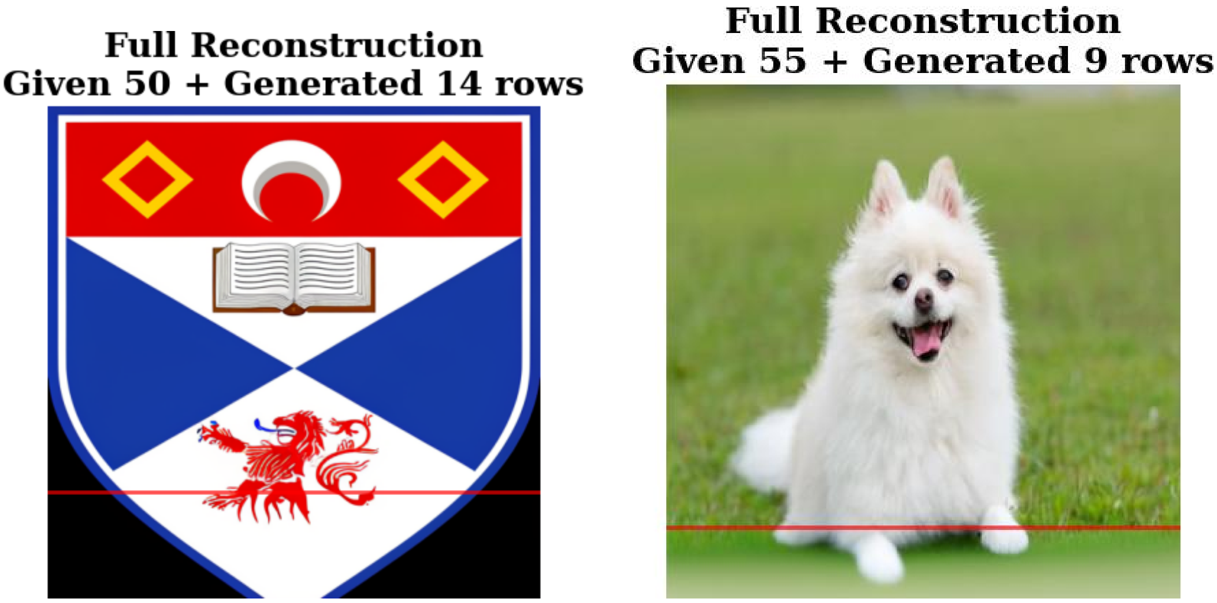

Figure 6: Image generation evaluation. The model is conditioned on the subset of image tokens above the red line and autoregressively generates the remaining tokens to reconstruct the full image.

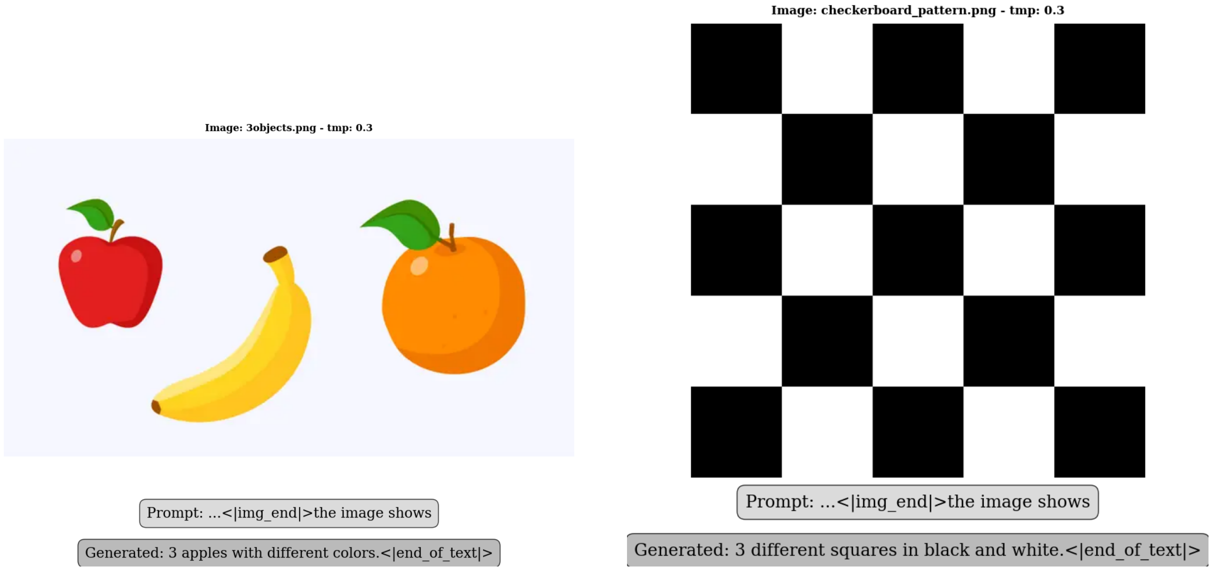
Image: 3objects.png - tmp: 0.3
Image: checkerboard_pattern.png - tmp: 0.3

Prompt: ...<|img_end|>the image shows

Generated: 3 apples with different colors.<|end_of_text|>

Prompt: ...<|img_end|>the image shows

Generated: 3 different squares in black and white.<|end_of_text|>

Figure 7: Caption generation evaluation. The model is conditioned on the full set of image tokens and the sentence prefix "the image shows". It then autoregressively generates the remaining text tokens as a caption.