

## Appendix

### Reconstruction of the various encoders

As an ablation study, we train the reconstructor for several of the most well-known vision encoders that are widely used in computer vision applications. Specifically, we select multiple families of ViT-based encoders that vary in parameter count, architectural details, training objectives, and pretraining datasets. In Tab. 1, we present the architectural analysis of the evaluated vision encoders. Tab. 3 and 4 show training peculiarities of the analyzed encoders.

### Additional visualizations

Fig. 2 shows additional examples of changing the anal with myestas, Fig. 3 shows additional examples of suppressing the blue channel, and Fig. 4 examples for colorization.

### Training Details

Reconstructors were trained with Adam and a cyclic learning-rate schedule. Hyperparameters are listed in Tab. 2. Training on 3xA100 (80 GB) GPUs took 6–24 h, depending on resolution.

### References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3
- [2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 2
- [3] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 2
- [4] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 2
- [5] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Jifeng Dai, and Wenhai Wang. Mini-intervl: A flexible-transfer pocket multimodal model with 52
- [6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

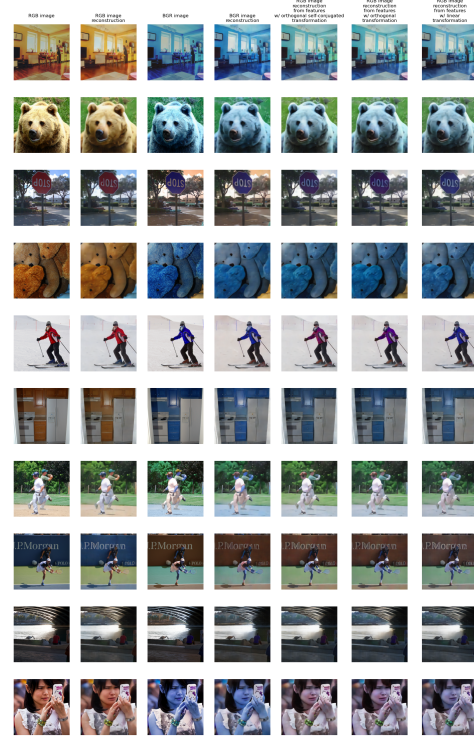


Figure 2. Color-swap via simple transformations in SigLIP2 feature space. Each row presents: (1) the original image, (2) its reconstruction from encoder features, (3) the image after swapping red and blue channels in pixel space, (4) the reconstruction of the pixel-swapped image, (5) the reconstruction obtained by applying the corresponding orthogonal self-conjugated channel-swap directly in feature space, (6) the reconstruction obtained by applying the corresponding orthogonal channel-swap directly in feature space, (7) the reconstruction obtained by applying the corresponding linear channel-swap directly in feature space. The near-identical results in columns 4 and 5, 6, 7 confirm that simple transformations in latent space induce coherent, interpretable color edits in the reconstructed images.

- [7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2
- [8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 2
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Table 1. Comparison of Image Encoders: input resolution, sequence length, parameter count, embedding dimension

Model	Resolution	Sequence length	#Params Vision tower	Out di- mension
<b>CLIP</b>				
timm/EVA-02 [4]	224×224 px	196 (14×14)	86 M	768
OpenAI CLIP [9]	224×224 px	196 (14×14)	86 M	768
LAION CLIP <sup>1</sup>	224×224 px	196 (14×14)	86 M	768
Facebook MetaCLIP [11]	224×224 px	196 (14×14)	86 M	768
Apple DFN2B-CLIP [3]	224×224 px	196 (14×14)	86 M	768
<b>SigLIP</b>				
SigLIP [12]	224×224 px	196 (14×14)	93 M	768
SigLIP [12]	256×256 px	256 (16×16)	93 M	768
SigLIP [12]	384×384 px	576 (24×24)	93 M	768
<b>SigLIP2</b>				
SigLIP2 [10]	224×224 px	196 (14×14)	93 M	768
SigLIP2 [10]	256×256 px	256 (16×16)	93 M	768
SigLIP2 [10]	384×384 px	576 (24×24)	93 M	768
SigLIP2 [10]	512×512 px	1024 (32×32)	93 M	768
<b>SAM</b>				
Facebook SAM [7]	1024×1024 px	4096 (64×64)	90 M	768
<b>DinoV2</b>				
DinoV2 [8]	518×518 px	1369 (37×37)	87M	768
<b>InternViT</b>				
InternViT-V1.5-300M [5]	448×448 px	1024 (32×32)	304 M	1024
InternViT-V2.5-300M [2]	448×448 px	1024 (32×32)	304 M	1024

Table 2. Reconstructor Training Hyperparameters

Hyperparameter	Value
Optimizer	Adam [6]
Learning Rate	3e-4
Learning Rate Scheduler	Cyclic
Adam Beta1	0.9
Adam Beta2	0.999
Batch Size (per device)	10
Training Epochs	40

- Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2
- [11] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2024. 2
- [12] 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. 2
- [13] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 3

075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091

- 069 Amanda Askill, 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. 2
- 070
- 071
- 072
- 073 [10] Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil
- 074

Table 3. Comparison of Image Encoders: training objective and architecture

Model	Training Objective
<b>CLIP</b>	
timmm/EVA-02	Contrastive image–text alignment. Weights initied from EVA model trained on Masked image modeling: reconstructing CLIP features from masked patches (negative cosine loss)
OpenAI CLIP	Contrastive image–text alignment (InfoNCE)
LAION CLIP	Contrastive on LAION-2B (2 B image–text pairs)
Facebook MetaCLIP	Contrastive on CommonCrawl 2.5 B data
Apple DFN2B-CLIP	Contrastive on DFN-2B filtered data
<b>SigLIP</b>	
SigLIP	Pairwise sigmoid contrastive loss
<b>SigLIP2</b>	
SigLIP2	Multitask: sigmoid contrastive, captioning, self-distillation, masked modeling
<b>SAM</b>	
Facebook SAM	Promptable segmentation: predict masks from sparse or dense prompts
<b>DinoV2</b>	
DinoV2	Discriminative self-supervised pretraining: self-distillation from Dino [1] and masked image modeling from iBOT [13]
<b>InternViT</b>	
InternViT-V1.5-300M	Contrastive pre-training, LLM alignment, distillation
InternViT-V2.5-300M	Contrastive pre-training, LLM alignment with progressive scaling, distillation

Table 4. Comparison of Image Models by Pretraining Objectives

Model	Contrastive	Captioning	Masked Modeling	Segmentation	LLM Align- ment	Self- Distillation
<b>CLIP</b>						
timmm/EVA-02	✓		✓			
OpenAI CLIP	✓					
LAION CLIP	✓					
Facebook	✓					
MetaCLIP						
Apple	✓					
DFN2B-CLIP						
<b>SigLIP</b>						
SigLIP	✓					
<b>SigLIP2</b>						
SigLIP2	✓	✓	✓			
<b>SAM</b>						
Facebook SAM				✓		
<b>DinoV2</b>						
DinoV2			✓			✓
<b>InternViT</b>						
InternViT-V1.5-300M	✓				✓	
InternViT-V2.5-300M	✓				✓	



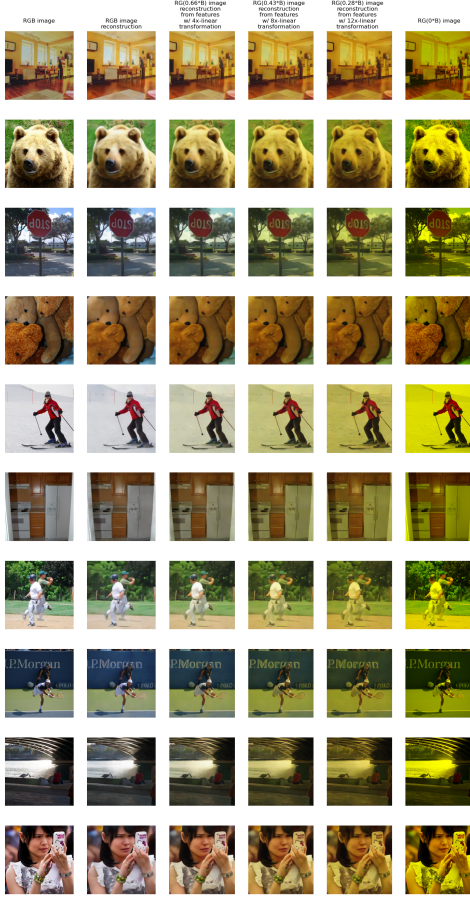


Figure 3. B-channel suppression via linear transformations in SigLIP2 feature space. Each row presents: (1) the original image, (2) its reconstruction from encoder features (3) reconstruction obtained by quadrupling the corresponding linear blue channel suppression operator directly in the feature space, (4) reconstruction obtained by applying the corresponding linear blue channel suppression operator eight times directly in the fisheye space, (5) reconstruction obtained by twelvefold the corresponding linear blue channel suppression operator directly in the feature space, (6) the image after blue channel nulling in pixel space. The near-identical results in columns 5 and 6 confirm that simple transformations in latent space induce coherent, interpretable color edits in the reconstructed images.

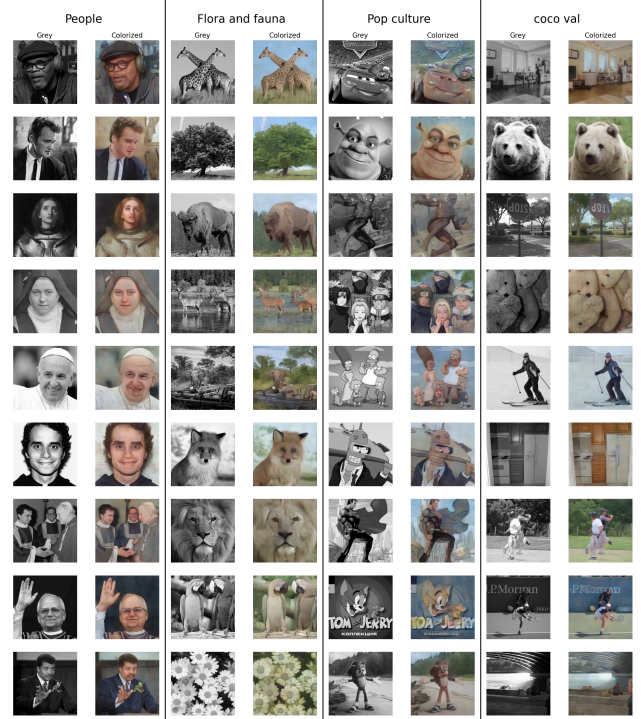


Figure 4. Examples of solving the colorization problem by applying a linear transformation in the feature space.