GLASS: UNLOCKING DENSE PERCEPTION IN GENERATIVE VISION ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative vision encoders pretrained with an autoregressive caption objective transfer well to multimodal LLMs, but their patch-level representations lag on dense perception. Caption prediction does not explicitly supervise the geometry or absolute position of individual patches. In our transfer experiments, existing dense post-training recipes improve dense prediction but reduce downstream captioning and multimodal understanding. We explore retaining the caption objective while adding masked feature recovery. Here the mask selects both the recovery targets and the visual evidence available for caption prediction. Hiding the patches receiving the most caption attention produces a failure pattern we term **text prior collapse**. Features from different images become more similar and dense performance declines, yet the caption loss remains close to the other masking runs and does not reflect the extent of the degradation. We propose **G**rounded **LA**yout and **S**emantic ma**S**king (**GLASS**), a caption-guided dense post-training recipe. Semantic masking keeps the highest-attention half of the patches visible, and masked feature recovery uses this context to predict full-image teacher features at hidden locations. Layout recovery adds explicit geometric supervision by re-encoding the visible patches without positional information and regressing their normalized coordinates. On GenLIP and OpenVision 2, GLASS improves dense perception and downstream captioning, with near-pretrained ImageNet accuracy and aggregate multimodal scores under the evaluated protocols. Further dense evaluations show gains across model scales.

1 INTRODUCTION

Generative pretraining encoders couple a vision encoder with a text decoder and train the resulting model with an autoregressive language-modeling objective (Fang et al., 2026; Fini et al., 2025; Liu et al., 2026). Compared with contrastive pretrained encoders, this objective is closer to the next-token prediction that a multimodal LLM is ultimately trained with, so the resulting features are more compatible with the generative objective of the downstream LLM (Fang et al., 2026). However, a correct caption does not require every patch to preserve object boundaries, local geometry, or absolute position, because the objective rewards whatever helps the caption as a whole rather than the geometry of individual patches. On dense perception tasks such as semantic segmentation and depth estimation, the patch-level spatial representations of these encoders therefore lag well behind their image-level competence.

A natural response is to continue training a pretrained encoder with local spatial objectives. Some CLIP-based recipes rely on a text tower or text-derived region supervision (Zhao et al., 2026; Zohra et al., 2025), whereas language-free methods such as NeCo (Pariza et al., 2025) and DIP (Sirko-Galouchenko et al., 2025) can be transferred directly to generative encoders. In our transfer experiments, these methods improve dense prediction but reduce downstream captioning and multimodal understanding (Figure 1 (c)). These results motivate a post-training recipe that strengthens spatial representations while retaining the capabilities learned through caption pretraining.

Masked prediction offers a natural way to add local supervision (He et al., 2022; Zhou et al., 2021). We explore this route by retaining the caption objective and adding masked feature recovery, in which a student predicts the full-image teacher's features at hidden patch locations. This provides patch-level targets without dense annotations. The masking rule, however, now serves two tasks. It

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

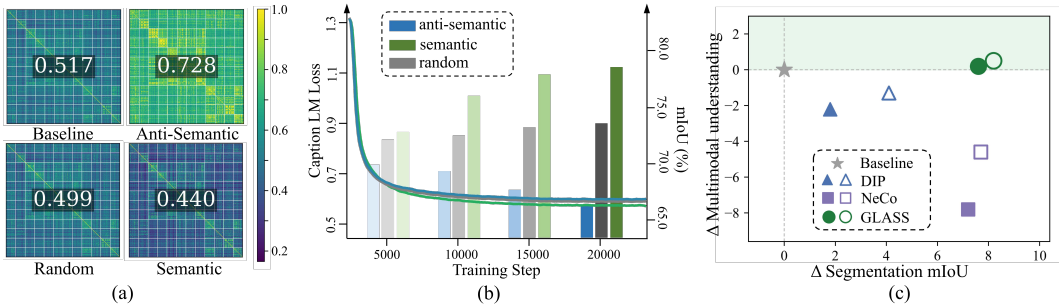

Figure 1: **Masking effects and post-training outcomes.** (a) Pairwise cosine similarity between image-level features in COCO for the pretrained encoder and the three masking rules on GenLIP-L. (b) The same runs during post-training, showing the caption LM loss and PASCAL-VOC segmentation mIoU. (c) Dense gains and changes in multimodal understanding relative to each pretrained encoder, with filled markers for GenLIP L/16 and open markers for OpenVision 2 L/14.

selects the locations to recover and determines the visual evidence available for caption prediction. A mask that makes recovery difficult may also remove the evidence needed to predict the caption.

In this joint training setup, we observe a failure pattern that we term **text prior collapse**. Hiding the patches receiving the most caption attention increases cross-image feature similarity (Figure 1 (a)) and degrades dense performance, whereas semantic and random masking improve dense performance. Yet the caption losses remain close (Figure 1 (b)), so caption loss alone does not capture the extent of visual degradation in these runs. An autoregressive decoder can predict many tokens from preceding text. Prior work reports increased language-prior reliance under visual corruption (Leng et al., 2024) and diminishing visual influence during generation (Hu et al., 2025). Consistent with this interpretation, our diagnostic shows reduced sensitivity to image–caption correspondence under anti-semantic masking, despite only a small change in correct-pair NLL (Appendix C.4). The contrast between masking rules motivates retaining caption-relevant evidence when adding masked recovery.

We propose **G**rounded **LA**yout and **S**emantic ma**S**king (**GLASS**) to combine evidence-preserving masking with local feature and geometric supervision. Semantic masking scores patches using caption-to-patch attention and keeps the highest-scoring half visible for caption prediction. Masked feature recovery uses these visible patches to recover teacher features on the hidden half. Recovering teacher features, however, does not explicitly require the representation to encode patch coordinates. Layout recovery therefore adds a geometric objective by re-encoding the visible semantic patches without positional information and regressing their normalized coordinates. Unlike a categorical grid index, the coordinate target penalizes nearby errors less than distant ones. GLASS thus uses the caption both as a retained training objective and as a guide to the visible patches from which feature and layout recovery are learned.

Our contributions are threefold.

- We introduce **GLASS**, a caption-guided dense post-training recipe that combines semantic masking, masked feature recovery, and layout recovery.
- We identify **text prior collapse** in masked post-training, where visual representations degrade without a commensurate change in caption loss, and show that the choice of visible patches substantially affects representation quality.
- We validate GLASS on GenLIP and OpenVision 2, showing improved dense perception and downstream captioning, with near-pretrained ImageNet accuracy and aggregate multimodal scores. Dense improvements persist across the evaluated model scales.

2 Related Work

Generative Pretraining Encoders. Unlike contrastive encoders that align image and text embeddings (Radford et al., 2021; Cherti et al., 2023; Zhai et al., 2023; Tschannen et al., 2025), generative

encoders learn visual representations through caption prediction. Early generative encoders such as SimVLM (Wang et al., 2021) and GIT (Wang et al., 2022) showed that caption generation yields transferable visual representations, CoCa (Yu et al., 2022) introduced contrastive captioning, and CapPa (Tschannen et al., 2023) scaled plain image captioners into vision learners. GenLIP (Fang et al., 2026), AIMv2 (Fini et al., 2025), and OpenVision 2 (Liu et al., 2026) further scale generative pretraining with strong multimodal transfer. Caption training can encode spatial information, but does not explicitly supervise each patch's geometry and coordinates. We study dense post-training that strengthens these representations while retaining caption-pretrained capabilities.

Language Priors and Modality Collapse. Strong autoregressive decoders can ignore conditioning representations, as observed for latent variables in text VAEs (Bowman et al., 2016). Related work documents language-prior reliance in image captioning (Rohrbach et al., 2018) and visual question answering (Agrawal et al., 2018). VCD shows that visual corruption can amplify language-prior reliance (Leng et al., 2024), while EVRB reports that visual-conditioned predictions can approach an estimated language-prior distribution during generation (Hu et al., 2025). Modality collapse has also been studied in multimodal fusion, where models learn to rely on only a subset of the input modalities (Chaudhuri et al., 2025). These studies provide context for conditional-input neglect. Our observation concerns masked post-training of the vision encoder, where cross-image feature similarity increases and dense performance declines despite caption losses remaining close across masking rules.

Informed Masking. Masking is a core mechanism of self-supervised learning (He et al., 2022). AttMask (Kakogeorgiou et al., 2022) and SdAE (Chen et al., 2022) mask highly attended regions to strengthen the learning signal, Attentive Mask CLIP (Yang et al., 2023) filters tokens to focus contrastive supervision on informative regions, MILAN (Hou et al., 2022) reconstructs CLIP features at CLIP-selected locations, and SemMAE (Li et al., 2022) and SemMIM (Liu et al., 2024b) guide mask placement with object or text semantics. Closest to our masking rule, SyCoCa (Ma et al., 2024) uses caption relevance to guide pixel reconstruction and captioning within contrastive-captioning pretraining. We instead study post-training of purely generative encoders, retaining caption prediction while adding feature and coordinate recovery. Our semantic masking selects the visible context for both caption prediction and recovery.

Dense Post-training. Dense post-training of vision encoders has been widely studied. CLIP-based dense adaptation often uses region–language supervision (Zhao et al., 2026; Zohra et al., 2025), while CLIPSelf (Wu et al., 2024) and ATAS (Yeo et al., 2025a) use language-free patch-level self-distillation. Other language-free approaches include NeCo's patch-ordering objective (Pariza et al., 2025), DIP's in-context segmentation with diffusion pseudo-masks (Sirko-Galouchenko et al., 2025), and PaKA's teacher–student patch-relation alignment (Yeo et al., 2025b). DIP is agnostic to the base encoder. We evaluate DIP and NeCo on generative encoders, including variants retaining the caption objective, to test whether spatial gains can be obtained while retaining downstream captioning and multimodal capabilities. GLASS combines caption-guided masking with feature and layout recovery toward this goal.

3 GLASS

3.1 Preliminaries

A generative vision encoder is trained to generate a caption for its input image. Its vision tower ϕ patchifies the image and maps the patches into embeddings $\mathbf{v}_{1:N} = \phi(x_{1:N})$, which condition the generation of the caption tokens $y_{1:S}$. Pretraining minimizes an autoregressive caption objective on image–caption pairs,

$$\mathcal{L}_{\mathrm{LM}}(\theta) \;=\; -\,\mathbb{E}_{(x,y)} \sum_{t=1}^{S} \log p_\theta\big(y_t \,\big|\, y_{<t},\, \mathbf{v}_{1:N}\big), \tag{1}$$

predicting each caption token from the preceding text and the visual embeddings. Although defined on text tokens, this loss also trains the vision encoder through the visual embeddings. We retain this objective during post-training.

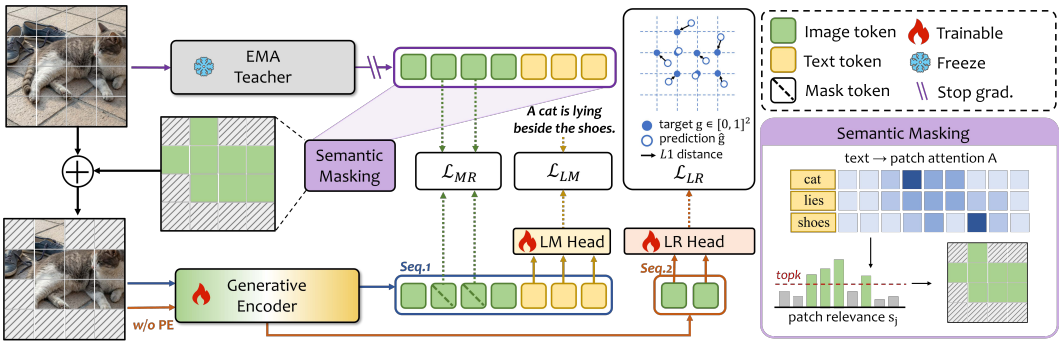

Figure 2: **Overview of GLASS.** The EMA teacher processes the complete image, scores patches by caption attention, and provides feature targets. The student uses the masked view for caption prediction (LM) and masked feature recovery (MR). A separate pass re-encodes the visible semantic patches without positional information for coordinate regression in layout recovery (LR).

GLASS post-trains such an encoder on image–caption pairs without dense annotations. An EMA teacher $f_{\bar{\theta}}$, initialized from the pretrained weights θ_0, tracks the student parameters with momentum α. As shown in Figure 2, the teacher processes the complete image and serves two roles. It scores each patch by how strongly the caption attends to it, which decides the mask, and its patch features provide the targets for masked recovery. The student f_θ processes a masked view of the same image and learns from the retained caption objective together with two dense terms,

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{MR}}\,\mathcal{L}_{\mathrm{MR}} + \lambda_{\mathrm{LR}}\,\mathcal{L}_{\mathrm{LR}}, \tag{2}$$

where $\mathcal{L}_{\mathrm{MR}}$ recovers masked features and $\mathcal{L}_{\mathrm{LR}}$ recovers layout, both defined below. The three objectives are optimized jointly through the shared vision backbone.

3.2 Semantic Masking

In joint caption prediction and masked recovery, the masking rule selects both the hidden recovery targets and the visible evidence. Figure 1 shows that hiding high-attention patches can degrade visual representations even when caption prediction is retained. This motivates keeping the patches ranked most relevant to the caption visible and using the remainder as recovery targets.

We score each patch by the attention it receives from the caption. Let $\mathcal{Q}$ collect the caption positions whose next-token target is a lexical token, so that non-lexical targets are excluded. In layer l of the last four teacher layers, a caption position $j \in \mathcal{Q}$ attends to patch i with the head-averaged weight

$$w_{ji}^{l} \;=\; \frac{1}{n_h}\sum_{h=1}^{n_h} \mathrm{softmax}_i\!\left(\frac{(\mathbf{q}_j^{l,h})^\top \mathbf{k}_i^{l,h}}{\sqrt{d}}\right), \tag{3}$$

where $\mathbf{q}_j^{l,h}$ and $\mathbf{k}_i^{l,h}$ are the query of caption position j and the key of patch i in head h. Averaging over the 4 layers and the caption positions yields the importance of patch i,

$$s^i = \frac{1}{4}\sum_{l}\frac{1}{|\mathcal{Q}|}\sum_{j\in\mathcal{Q}} w_{ji}^{l}. \tag{4}$$

Caption-to-patch attention provides a model-internal proxy for caption relevance. Averaging across heads and layers combines information from multiple attention maps rather than relying on a single map.

The top-scoring half of the patches forms the semantic set and the remainder forms the masked set,

$$\mathcal{P} = \mathrm{top}_{N/2}\big(\{s^i\}_{i=1}^{N}\big), \qquad \mathcal{M} = \{1,\dots,N\}\setminus\mathcal{P}, \tag{5}$$

where $\mathrm{top}_{N/2}$ returns the $N/2$ patches of highest score. Let $\mathbf{e}_i$ denote the input patch embedding before the transformer blocks of ϕ, distinct from the encoded feature $\mathbf{v}_i$. We keep the semantic patches visible and replace the remaining input embeddings with a learnable mask token $\mathbf{m}$,

$$\tilde{\mathbf{e}}_i = \begin{cases} \mathbf{m}, & i \in \mathcal{M}, \\ \mathbf{e}_i, & i \in \mathcal{P}. \end{cases} \tag{6}$$

The student transformer then processes $\tilde{\mathbf{e}}_{1:N}$ with the usual positional information for caption prediction and masked feature recovery. Thus, masking precedes contextual encoding. The visible set supplies caption-selected context for predicting the caption and recovering hidden features. The masked set contains lower-attention patches, which can still carry semantic and geometric information. Masked feature recovery supervises these hidden locations, while layout recovery re-encodes the visible set to predict its coordinates. Masking visualizations are provided in Appendix E.

3.3 Masked Feature Recovery

Masked feature recovery reconstructs the hidden half of the image at the feature level. For each masked position, the student feature passes through a projection head and is matched against the teacher feature at the same location in a self-distillation form. Let g_S and g_T be the student and teacher projection heads and $\mathbf{c}$ a running mean updated by EMA.

$$\mathcal{L}_{\mathrm{MR}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \langle \mathbf{p}_i^T, \log \mathbf{p}_i^S \rangle. \tag{7}$$

The two distributions are

$$\mathbf{p}_i^S = \mathrm{softmax}\Big(\tfrac{g_S(\mathbf{h}_i)}{\tau_S}\Big), \qquad \mathbf{p}_i^T = \mathrm{softmax}\Big(\tfrac{g_T(\mathrm{sg}(\mathbf{v}_i^T)) - \mathbf{c}}{\tau_T}\Big), \tag{8}$$

where $\mathbf{h}_i$ is the student feature at the masked position, $\mathbf{v}_i^T$ its teacher counterpart, and τ_S, τ_T the temperatures. The running mean $\mathbf{c}$ tracks the teacher outputs with momentum μ. The projection heads and teacher-target centering and sharpening follow Zhou et al. (2021); Caron et al. (2021).

We formulate the task at the feature level rather than the pixel level (Ma et al., 2024) because visual features are the input to both caption prediction and downstream dense heads. The teacher sees the complete image, whereas the student must predict its targets from an incomplete view. This encourages the student to use relationships between visible content and hidden regions, supplying patch-level supervision without dense annotations. Semantic masking determines the context available for this recovery task while retaining high-attention patches for caption prediction.

3.4 Layout Recovery

Neither caption prediction nor feature recovery explicitly requires each patch representation to predict its coordinates. Layout recovery adds this geometric target to complement masked feature recovery. Following work on position recovery (Doersch et al., 2015; Noroozi & Favaro, 2016; Chen et al., 2023; Zhai et al., 2022; Wang et al., 2023; Caron et al., 2024), we train the encoder to infer the coordinates of the semantic patches without positional input. Writing $\phi(x; \mathbf{p})$ for the tower applied to patches x under a position assignment $\mathbf{p}$, we re-encode the semantic patches with position nulled to obtain position-free features $\tilde{\mathbf{h}}_{1:|\mathcal{P}|} = \phi(x_{\mathcal{P}}; \mathbf{0})$, where $\mathbf{0}$ assigns every patch the same null position. A rotary tower interprets $\mathbf{0}$ as a zero rotation angle, so its rotation collapses to the identity ($\cos 0 = 1$, $\sin 0 = 0$), whereas an absolute tower injects no position embedding. A lightweight head ψ then regresses the normalized coordinates of each patch, $\hat{\mathbf{g}}_i = \psi(\tilde{\mathbf{h}}_i)$, trained with

$$\mathcal{L}_{\mathrm{LR}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \big\| \hat{\mathbf{g}}_i - \mathbf{g}_i \big\|_1, \qquad \mathbf{g}_i = \Big(\frac{c_i}{W-1}, \frac{r_i}{H-1} \Big), \tag{9}$$

where (r_i, c_i) is the row and column of patch i in the $H \times W$ grid and $\hat{\mathbf{g}}_i$ is the two-dimensional prediction. Removing the positional input makes coordinate prediction depend on content cues from the visible patches rather than a supplied position embedding.

We formulate position recovery as coordinate regression rather than classification over grid indices. Unlike a categorical target, the coordinate loss explicitly encodes spatial distance. Under the L1 objective, predictions with smaller Manhattan distance to the true location incur smaller penalties. The layout branch uses the semantic set, tying coordinate prediction to the content retained for captioning. The ablations in Section 4.3 test both the contribution of this geometric objective and the choice of semantic rather than randomly selected patches as its input.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. We evaluate GenLIP (Fang et al., 2026) and OpenVision 2 (Liu et al., 2026), and report ablations on GenLIP-L/16. Every post-training run starts from the officially released checkpoint of the corresponding model. GLASS post-training uses an input resolution of 224×224 and runs on Recap-DataComp-12M (Li et al., 2024c) for 1 epoch on 16 NVIDIA L20 GPUs. All GLASS runs share the same optimizer configuration, with a backbone learning rate of 5e-6 and a learning rate of 1e-4 for the lightweight layout recovery head. Unless noted otherwise, we mask the bottom-scoring half of the patches and set the objective weights to λ_{MR}=0.3 and λ_{LR}=0.1. Further training details are provided in Appendix A and the complete evaluation protocols in Appendix B.

Semantic segmentation directly probes the quality of patch-level features. Following Siméoni et al. (2025), we train a linear segmentation head on top of the frozen encoder without fine-tuning of the backbone, and evaluate on PASCAL-VOC (Everingham et al., 2010), ADE20K (Zhou et al., 2017), and Cityscapes (Cordts et al., 2016). We report mean Intersection-over-Union (mIoU) on each validation set. We additionally evaluate training-free nearest-neighbor segmentation on PASCAL-VOC and ADE20K following DIP (Sirko-Galouchenko et al., 2025) (Appendix C.2).

Monocular depth estimation measures how well patch features encode local geometry. Following Siméoni et al. (2025), we train a linear depth probe on top of the frozen encoder and report RMSE on NYUv2 (Silberman et al., 2012) and the Eigen split of KITTI (Geiger et al., 2013).

Image classification assesses how dense post-training affects global recognition. Following the GenLIP evaluation protocol (Fang et al., 2026), we report attentive-probe top-1 accuracy on ImageNet-1K (Russakovsky et al., 2015).

Multimodal understanding tests whether the encoder's features support downstream reasoning in a multimodal LLM. We attach the frozen encoder to LLaVA-NeXT-Qwen2.5-1.5B (Liu et al., 2024a; Yang et al., 2025) without further training the visual tower, and report each benchmark's official accuracy metric across document and text understanding (ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), OCR-Bench (Liu et al., 2024c), SEED-Bench-2-Plus (Li et al., 2024b)) and science question answering (ScienceQA (Lu et al., 2022)), together with the standard score of the MME perceptual subset (Fu et al., 2025), reported out of its 2000-point maximum.

Image captioning evaluates the encoder's utility for downstream caption generation. With the same frozen-encoder VLM, we report CIDEr on TextCaps (Sidorov et al., 2020) and NoCaps (Agrawal et al., 2019).

Compared methods. We compare GLASS with DIP (Sirko-Galouchenko et al., 2025) and NeCo (Pariza et al., 2025), starting from the same released GenLIP and OpenVision 2 checkpoints and using common downstream evaluation protocols. For DIP and NeCo, we follow their published post-training recipes on COCO. The default DIP recipe updates only the last three backbone blocks; the full-backbone comparison is reported in Appendix C.3. We additionally evaluate DIP and NeCo with the caption objective retained on GenLIP-L/16. These comparisons evaluate complete recipes with different training data and budgets, detailed in Appendix C.1, rather than isolating objectives under matched data and compute. We also report the contrastive SigLIP 2 (Tschannen et al., 2025), the region-aware PE-core (Bolya et al., 2025), and the self-supervised Franca (Venkataramanan et al., 2026) and Web-DINO (Fan et al., 2025) as reference encoders. Their different architectures and pretraining corpora make them context for dense representation quality rather than controlled post-training comparisons.

4.2 RESULTS

Dense and global perception. Table 1 compares the dense and global results on GenLIP and OpenVision 2. Within the two generative-encoder blocks, GLASS ranks first in nine of the twelve entries and second in the remaining three. It raises segmentation mIoU by **6.2** to **9.8** points and lowers depth RMSE by **7.8%** to **26.7%** relative to the pretrained encoders, while ImageNet accuracy stays within 0.2 points on both architectures. GenLIP-L/16 with GLASS also exceeds all listed

Table 1: **Evaluation for dense and global prediction tasks.** Rows highlighted with and denote GLASS and reference pretrained encoders, respectively. **Bold** and underlined values indicate the best and second-best results within each generative-encoder block, respectively.

Method	Arch	↑ Segmentation			↓ Depth		↑ ImageNet classif.
		PASCAL	ADE20K	Cityscapes	NYUv2	KITTI	Attentive
SigLIP 2	g/16	72.7	42.7	64.8	0.494	3.273	88.0
PE-core	G/14	69.2	38.9	61.1	0.590	4.119	–
Franca	g/14	82.9	46.3	68.7	0.445	3.140	–
Web-DINO	7B/14	76.1	42.7	68.3	0.466	3.158	–
GenLIP	L/16	70.8	41.0	61.6	0.507	3.554	82.8
+ DIP	L/16	73.6	43.2	62.0	0.506	3.514	82.8
+ NeCo	L/16	**80.5**	46.8	67.8	0.440	3.499	81.4
+ GLASS	L/16	78.9	**47.2**	**70.1**	**0.438**	**3.277**	**82.9**
OpenVision 2	L/14	68.9	37.5	60.5	0.599	4.374	**86.1**
+ DIP	L/14	74.1	41.3	63.9	0.579	4.202	85.9
+ NeCo	L/14	**80.2**	45.1	64.6	0.454	3.742	84.2
+ GLASS	L/14	78.7	**45.9**	**66.8**	**0.439**	**3.668**	85.9

Table 2: **Multimodal understanding and captioning under LLaVA-NeXT-Qwen2.5-1.5B.** The encoder is frozen while the projector and the language model are fine-tuned. Δ reports the mean score change over the metrics of the same group with respect to the pretrained encoder of the block, with MME-P first rescaled to a percentage of its 2000 maximum so that it is commensurate with the other multimodal scores, following Fang et al. (2026).

Model	Arch	↑ Multimodal understanding							↑ Captioning		
		ChartQA	TextVQA	OCR-B	SEED-2	MME-P	SQA	Δ	TextCaps	NoCaps	Δ
GenLIP	L/16	41.2	53.6	**51.1**	**51.1**	1258	**76.1**	–	131.4	82.6	–
+ DIP	L/16	39.3	50.5	49.8	47.5	1248	73.4	↓2.2	124.4	78.8	↓5.4
+ NeCo	L/16	34.4	45.6	45.5	42.8	1175	62.1	↓7.8	118.9	72.3	↓11.4
+ GLASS	L/16	**42.4**	**54.2**	51.0	50.3	**1264**	**76.1**	↑0.2	**135.4**	**84.1**	↑2.8
OpenVision 2	L/14	24.1	35.8	33.1	44.6	1242	77.0	–	107.0	83.5	–
+ DIP	L/14	23.2	32.6	29.8	**45.2**	**1248**	76.0	↓1.3	101.4	82.5	↓3.3
+ NeCo	L/14	21.4	29.1	27.9	41.0	1156	72.1	↓4.6	99.2	81.2	↓5.1
+ GLASS	L/14	**25.8**	**36.0**	**33.6**	44.2	**1248**	**77.8**	↑0.5	**108.5**	**85.2**	↑1.6

reference encoders on ADE20K and Cityscapes, reaching 47.2 and 70.1 compared with Franca's 46.3 and 68.7.

The recipes differ in the balance they achieve. DIP changes ImageNet accuracy little, but its depth gains on GenLIP are small (NYUv2 -0.001, KITTI -0.040). NeCo exceeds GLASS on PASCAL by 1.6 and 1.5 mIoU points, but lowers ImageNet accuracy by 1.4 and 1.9 points relative to pretraining. GLASS combines improvements in segmentation and depth with near-pretrained ImageNet accuracy. The dense gains also extend beyond linear probing: GLASS improves over the pretrained encoders in training-free retrieval segmentation on both PASCAL-VOC and ADE20K across all evaluated support fractions (Appendix C.2).

Multimodal understanding and captioning. Table 2 evaluates each frozen encoder in LLaVA-NeXT-Qwen2.5-1.5B (Liu et al., 2024a; Yang et al., 2025). GLASS's aggregate multimodal scores remain close to pretraining, with mean changes of $+0.2$ on GenLIP and $+0.5$ on OpenVision 2. This average includes task-level decreases: SEED-2 falls by 0.8 and 0.4 points, respectively. Both captioning benchmarks improve on both architectures, averaging $+2.8$ and $+1.6$ CIDEr; on GenLIP, TextCaps rises from 131.4 to **135.4** and NoCaps from 82.6 to **84.1**.

The transferred baselines show larger aggregate decreases. NeCo lowers mean multimodal performance by 7.8 points on GenLIP and 4.6 on OpenVision 2, and mean captioning CIDEr by 11.4 and 5.1. DIP shows smaller decreases, with multimodal changes of $-2.2/-1.3$ and captioning changes of $-5.4/-3.3$. Together with Table 1, these results support GLASS's balance of dense gains and downstream language-related performance under the evaluated recipes.

Table 3: **Retaining the caption objective in dense post-training.** Results on GenLIP-L/16 with and without the LM objective. *Pretrained* denotes the original GenLIP-L/16 checkpoint before dense post-training, evaluated using the same downstream protocols.

Recipe	Dense task		Multimodal understanding				Captioning	
	PASCAL ↑	NYUv2 ↓	ChartQA ↑	OCR-B ↑	MME-P ↑	SQA ↑	TextCaps ↑	NoCaps ↑
Pretrained	70.8	0.507	41.2	**51.1**	1258	**76.1**	131.4	82.6
DIP	73.6	0.506	39.3	49.8	1248	73.4	124.4	78.8
NeCo	**80.5**	0.440	34.4	45.5	1175	62.1	118.9	72.3
DIP + LM	73.1	0.508	40.4	50.4	1249	75.1	126.0	79.2
NeCo + LM	77.3	0.465	37.3	48.1	1206	66.7	124.4	76.3
GLASS	78.9	**0.438**	**42.4**	51.0	**1264**	**76.1**	**135.4**	**84.1**

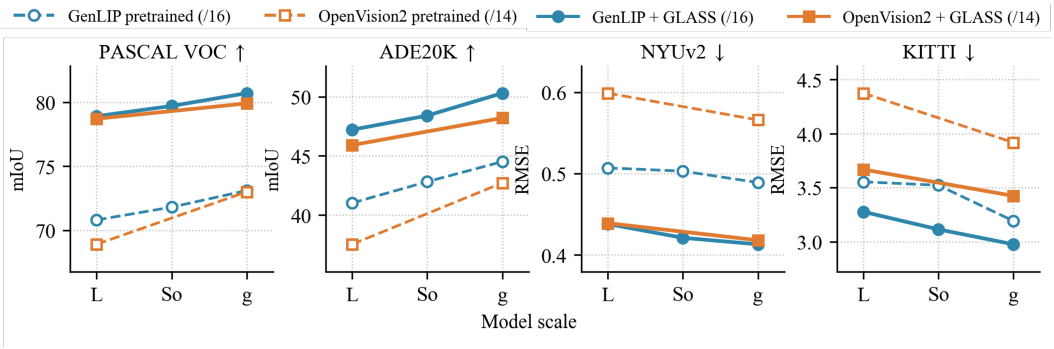

Figure 3: **Scaling behaviour of GLASS on dense perception.** GLASS raises segmentation mIoU and lowers depth RMSE at every scale for both architectures.

Baselines with the caption objective. Table 3 examines the effect of retaining the caption objective in the DIP and NeCo recipes. Adding LM improves both captioning benchmarks and all four multimodal scores for both methods, but neither recovers the pretrained scores on these six metrics. For NeCo, this partial recovery accompanies lower PASCAL mIoU (80.5 to 77.3) and higher NYUv2 RMSE (0.440 to 0.465); DIP's dense results change only slightly. This pattern suggests that adding caption supervision changes the balance of representation learning. The LM objective may shift encoder updates toward caption-relevant features, partially reducing the spatial specialization encouraged by NeCo's patch-ordering objective. GLASS achieves higher PASCAL mIoU and lower NYUv2 RMSE than both LM-augmented baselines, while also scoring higher on all four multimodal and both captioning benchmarks.

Scaling behavior across model sizes. Figure 3 extends the dense evaluation from L to So and g for GenLIP, and from L to g for OpenVision 2. GLASS improves segmentation and depth at every evaluated scale in both families. The persistent gains show that the benefit of post-training is not confined to the L models used in the main comparisons.

4.3 Ablation Studies

Loss components. Rows 2–5 of Table 4 examine the contributions of MR and LR with LM retained. Semantic-masked LM alone changes dense performance little from pretraining. Relative to this LM-only run, MR raises PASCAL mIoU by 4.9 points and lowers NYUv2 RMSE by 0.008, whereas LR yields changes of $+2.7$ and -0.044. MR thus provides the larger segmentation gain, while LR provides the larger reduction in depth error. MR asks the student to recover contextual teacher descriptors at hidden locations, which may strengthen the patch-level semantic discrimination needed for segmentation. LR instead predicts coordinates without positional input, encouraging the encoder to infer spatial layout from visual content; this may strengthen layout cues useful for monocular depth estimation. Combining MR and LR reaches 78.9 mIoU and 0.438 RMSE, the best results on both dense tasks in this ablation, supporting their complementary contributions while maintaining near-pretrained ImageNet accuracy.

Table 4: **Component and masking ablations.** Rows 2–5 ablate MR and LR with LM retained; row 5 is GLASS. Rows 5–7 compare masking rules with all three objectives retained.

Index	Mask	LM	MR	LR	Dense task		Global task	Multimodal understanding			
					PASCAL ↑	NYUv2 ↓	IN Classif. ↑	ChartQA ↑	OCR-B ↑	MME-P ↑	SQA ↑
1	*Pretrained*				70.8	0.507	82.8	41.2	51.1	1258	**76.1**
2	Semantic	✓			71.4	0.505	82.8	42.0	51.2	1261	76.0
3	Semantic	✓	✓		76.3	0.497	**82.9**	42.2	51.1	1262	75.8
4	Semantic	✓		✓	74.1	0.461	82.6	41.0	50.9	1240	75.8
5	Semantic	✓	✓	✓	**78.9**	**0.438**	**82.9**	**42.4**	51.0	**1264**	**76.1**
6	Random	✓	✓	✓	73.5	0.472	82.6	41.6	**51.3**	1255	76.0
7	Anti-sem.	✓	✓	✓	66.0	0.521	81.1	37.5	49.3	1210	71.3

Table 5: **Design-choice ablations.** (a) Categorical absolute-order target versus our coordinate regression for layout recovery. (b) Ratio of patches kept visible under semantic masking. (c) Input to the layout branch, either the semantic set shared with caption view or a freshly sampled random set.

Position	PASCAL ↑	NYUv2 ↓
Categorical	77.6	0.467
Coordinates	**78.9**	**0.438**

(a)

Visible	PASCAL ↑	ImageNet ↑
25%	**79.3**	81.6
50%	78.9	**82.9**
75%	76.8	**82.9**

(b)

Layout input	PASCAL ↑	NYUv2 ↓
Random set	76.3	0.472
Semantic set	**78.9**	**0.438**

(c)

Masking strategy. Rows 5–7 of Table 4 keep all three objectives and vary the masking rule. Random masking selects half of the patches uniformly, whereas anti-semantic masking hides the half receiving the most caption attention. Relative to semantic masking, random masking reduces PASCAL mIoU from 78.9 to 73.5 and increases NYUv2 RMSE from 0.438 to 0.472. Anti-semantic masking is worse, reaching 66.0 and 0.521, below the pretrained encoder on both tasks. Its caption loss gives no warning of the extent of this degradation, remaining close to the other runs even as its dense performance declines through training (Figure 1). This is the text prior collapse pattern discussed in Section 1. The comparison shows that retaining the caption objective does not by itself prevent visual degradation, and supports selecting visible evidence using caption attention.

Design choices. Table 5 examines three choices. (a) Position target. Replacing coordinate regression with categorical absolute-order prediction reduces PASCAL mIoU by 1.3 points and increases NYUv2 RMSE by 0.029, supporting a target that reflects spatial distance. Categorical targets treat grid locations as separate classes without explicitly encoding their proximity, whereas the L1 coordinate loss assigns smaller penalties to spatially closer predictions. This distance-aware supervision may help the encoder learn spatial structure useful for dense prediction. (b) Visible ratio. Keeping 25% of patches visible slightly improves PASCAL mIoU to 79.3, but lowers ImageNet accuracy to 81.6. Keeping 75% preserves ImageNet accuracy at 82.9 but reduces PASCAL mIoU to 76.8. The default 50% split balances these two outcomes. (c) Layout input. Using a freshly sampled random half of the patches instead of the semantic set reduces PASCAL mIoU by 2.6 points and increases NYUv2 RMSE by 0.034. Without positional input, the layout branch must infer coordinates from patch content. Caption-relevant patches may provide informative semantic cues for this inference, while sharing the semantic set directs caption and layout supervision to the same visible content.

5 Conclusion

We presented **GLASS**, a caption-guided post-training recipe for improving dense perception while retaining the capabilities of generative vision encoders. GLASS keeps the caption objective and uses caption attention to select the visible context for masked feature recovery and coordinate-based layout recovery. Our masking experiments identify text prior collapse, in which visual features deteriorate while the caption loss gives no warning of the extent of the degradation. This observation motivates considering the evidence available for caption prediction when designing masked post-training. Evaluations on GenLIP and OpenVision 2 show improved segmentation, depth estimation, and downstream captioning, with global recognition and aggregate multimodal understanding preserved. Dense gains also persist across the evaluated model scales.

AI USE STATEMENT

In this work, we have not used generative AI tools for any task in the required-disclosure category. Additionally, we used them for three tasks in the recommended-disclosure category: identifying and summarizing relevant literature; editing the manuscript to improve readability, spelling, and grammar; and creating or editing portions of the evaluation code. We have reviewed all AI-assisted work. Every reference surfaced by the literature search was verified against the original publication before being cited, and all AI-assisted code was manually inspected and validated by the authors before contributing to any reported result. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The formal definitions of the three GLASS components, semantic masking, masked feature recovery, and layout recovery, together with the post-training objective, are given in Section 3. Section 4.1 summarizes the experimental setup. Appendix A provides the post-training hyperparameters and implementation details, while Appendix B specifies the downstream evaluation protocols. Training budgets and computational costs are reported in Appendix C.1. All experiments rely exclusively on publicly available datasets. We will release the complete source code upon paper acceptance.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956, 2019. doi: 10.1109/ICCV.2019.00904.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 60884–60937. Curran Associates, Inc., 2025. doi: 10.52202/085713-2036. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/57bc0a850255e2041341bf74c7e2b9fa-Paper-Conference.pdf.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pp. 10–21, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers for semantic segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 116–126, 2024. doi: 10.1109/WACV57701.2024.00019.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7555–7577. PMLR, 2025.

Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European conference on computer vision*, pp. 108–124. Springer, 2022.

Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–13. IEEE, 2025.

Yan Fang, Mengcheng Lan, Zilong Huang, Weixian Lei, Yunqing Zhao, Yujie Zhong, Yingchen Yu, Qi She, Yao Zhao, and Yunchao Wei. Let vit speak: Generative language-image pre-training. *arXiv preprint arXiv:2605.00809*, 2026.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9641–9654. IEEE, 2025.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022.

Nanxing Hu, Xiaoyue Duan, Jinchao Zhang, and Guoliang Kang. Enhancing visual reliance in text generation: A bayesian perspective on mitigating hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4778–4787, 2025.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2022.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024b.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024c.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Haowei Liu, Yaya Shi, Haiyang Xu, Chunfeng Yuan, Qinghao Ye, Chenliang Li, Ming Yan, Ji Zhang, Fei Huang, Bing Li, et al. Semantics-enhanced cross-modal masked image modeling for vision-language pre-training. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14664–14675, 2024b.

Yanqing Liu, Xianhang Li, Letian Zhang, Zirui Wang, Zeyu Zheng, Yuyin Zhou, and Cihang Xie. Openvision 2: A family of generative pretrained visual encoders for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39164–39174, 2026.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Ziping Ma, Furong Xu, Jian Liu, Ming Yang, and Qingpei Guo. Sycoca: symmetrizing contrastive captioners with attentive masking for multimodal alignment. *arXiv preprint arXiv:2401.02137*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Valentinos Pariza, Mohammadreza Salehi, Gertjan J Burghouts, Francesco Locatello, and Yuki Asano. Near, far: Patch-ordering enhances vision foundation models' scene understanding. In *International Conference on Learning Representations*, volume 2025, pp. 101552–101579, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pp. 742–758. Springer, 2020.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Sophia Sirko-Galouchenko, Spyros Gidaris, Antonin Vobecky, Andrei Bursuc, and Nicolas Thome. Dip: Unsupervised dense in-context post-training of visual representations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4264–4274. IEEE, 2025.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil 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.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10533–10544, 2026.

Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tong Wang, and ZHAO-XIANG ZHANG. Droppos: Pre-training vision transformers by reconstructing dropped positions. *Advances in Neural Information Processing Systems*, 36:46134–46151, 2023.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International conference on learning representations*, volume 2024, pp. 52484–52503, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2759–2769. IEEE, 2023.

Juan Yeo, Soonwoo Cha, Jiwoo Song, Hyunbin Jin, and Taesup Kim. Atas: any-to-any self-distillation for enhanced open-vocabulary dense prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20390–20400. IEEE, 2025a.

Juan Yeo, Ijun Jang, and Taesup Kim. Patch-level kernel alignment for dense self-supervised learning. *arXiv preprint arXiv:2509.05606*, 2025b.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Yitan Cheng, Walter Talbott, Chen Huang, Hanlin Goh, and Joshua Susskind. Position prediction as an effective pretraining strategy. *arXiv preprint arXiv:2207.07611*, 2022.

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

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Chenyang Zhao, Wei Lin, Antoni B Chan, and Janet H Hsiao. Fine-grained clip fine-tuning with self-annotated region alignment. *arXiv preprint arXiv:2607.13661*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. β-clip: Text-conditioned contrastive learning for multi-granular vision-language alignment. *arXiv preprint arXiv:2512.12678*, 2025.

A Implementation Details

This appendix completes the post-training configuration summarized in Section 4.1. Table 6 lists the post-training hyperparameters shared by the GenLIP and OpenVision 2 runs. In OpenVision 2, the text decoder is updated jointly with the vision encoder during GLASS post-training. The input dimension D of both recovery heads matches the backbone hidden width, with D=1024, 1152, and 1536 for the L, So, and g models, respectively.

Table 6: **GLASS post-training hyperparameters.**

Hyperparameter	Value
Post-training data	Recap-DataComp-12M
Epochs	1
Input resolution	224×224
Backbone learning rate	5e-6
Layout-head learning rate	1e-4
Objective weights	λ_{MR}=0.3, λ_{LR}=0.1
Masking ratio	bottom-scoring 50%
Hardware	16× NVIDIA L20
Optimizer	AdamW
Global batch size	512
LR warmup ratio	0.05
LR decay style	cosine
LR decay ratio	1.0
LR min	1e-7
EMA teacher momentum α	0.996
Gradient clipping	1.0
Centering momentum	0.9
Teacher temperature	0.04
Student temperature	0.1

Semantic masking. Semantic masking is fully specified in Section 3, where the caption-to-patch attention is read from the last four teacher layers at every model size.

Masked feature recovery. The softmax cross-entropy loss, its centering, and the two temperatures are defined in Section 3. We specify here the projection heads and the features they read. The student head g_S and the teacher head g_T share a DINO-style architecture (Caron et al., 2021), namely a three-layer MLP with GELU activations ($D \to 4096 \to 4096 \to 256$), then LayerNorm, ℓ_2 normalization, and a bias-free linear layer to K=8192 prototypes. Both heads read the patch features from the last transformer block, after the final LayerNorm and before the text-path projection, the student at its masked positions and the EMA teacher at the same positions on the clean image.

Layout recovery. The L1 objective and the normalized coordinate target are defined in Section 3; the position-free re-encoding that produces the input features is detailed below. The head ψ is a three-layer MLP with ReLU activations that maps each D-dimensional backbone feature to its two-dimensional coordinate; its L-model configuration is $1024 \to 1024 \to 1024 \to 2$. Layout recovery regresses each visible patch's absolute position from a position-free re-encoding, so the tower must not leak position into that pass, and the two backbones suppress it differently. GenLIP uses rotary (MRoPE) embeddings, which cannot be skipped, so we cancel them by passing zero position ids on all three axes, where $\cos 0 = 1$ and $\sin 0 = 0$ reduce the rotation to the identity, and, critically, by setting `image_grid_thw=None`; otherwise the tower rebuilds the true coordinates from the grid and position leaks back in. OpenVision 2 uses absolute position embeddings, which we simply omit on this pass.

Baselines with the caption objective. The DIP + LM and NeCo + LM variants in Table 3 use the original COCO captions paired with their training images. The caption LM loss is added to the corresponding dense post-training objective with a coefficient of 1.0.

B Evaluation Protocols

Dense probes. For semantic segmentation (PASCAL-VOC, ADE20K, Cityscapes) and monocular depth (NYUv2, KITTI) we freeze the encoder and train only a linear probe, following the DINOv3 protocol (Siméoni et al., 2025). For segmentation the probe is a single 1×1 convolution (batch norm, dropout 0.1) on the last-layer patch tokens, trained with AdamW (learning rate 10^{-3}, weight

decay 10^{-3}) for $40{,}000$ iterations at batch 2 per GPU under the DINOv3 augmentation; inference uses a 512-pixel sliding window (stride 341), at a 512 short side for VOC and ADE20K and native 2048×1024 for Cityscapes, and we report mIoU over 21, 150, and 19 classes respectively. For depth the probe is a single 1×1 convolution into 256 linear bins whose expectation gives the predicted depth, trained with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}) for $38{,}400$ iterations under the DINOv3 depth loss with horizontal-flip test-time augmentation; we report RMSE in meters on metric depth without scale-shift alignment, within the Eigen crop on NYUv2 and the Garg crop on KITTI.

Global recognition. We report ImageNet-1K top-1 accuracy with an attentive probe following the GenLIP protocol (Fang et al., 2026). The probe pools the last-layer patch tokens with a single learnable query through 8-head cross-attention, followed by a layer norm and a linear classifier to 1000 classes. It is trained with AdamW (learning rate 3×10^{-4}, no weight decay) under a cosine schedule for 30 epochs at batch 128 per GPU, using random resized crops and horizontal flips at resolution 224; evaluation resizes to 256 and center-crops to 224.

Multimodal understanding and captioning. Following GenLIP (Fang et al., 2026), we attach the frozen encoder to LLaVA-NeXT (Liu et al., 2024a) through a two-layer MLP projector (Linear-GELU-Linear) and fully fine-tune the projector together with the Qwen2.5-1.5B-Instruct (Yang et al., 2025) language model, without LoRA or separate projector pretraining; the encoder supplies its last-layer normalized patch features rather than its original projection head and stays frozen. This supervised adaptation runs for one epoch on the LLaVA-OneVision instruction data (Li et al., 2024a) (global batch 128, AdamW, learning rate 10^{-5}, cosine schedule with 3% warmup) and serves only to build the downstream evaluation model, separate from our self-supervised post-training. GenLIP images are padded to a square and resized to 384 without AnyRes tiling, giving $(384/16)^2{=}576$ visual tokens, whereas OpenVision 2 uses 224×224 inputs, giving $(224/14)^2{=}256$ visual tokens. We evaluate with lmms-eval (Zhang et al., 2025) and its llava_onevision adapter under the qwen_1_5 template in the zero-shot setting, and report the official metric of each benchmark on its standard split (validation for TextVQA), namely relaxed accuracy for ChartQA, soft VQA accuracy for TextVQA, the OCR-Bench score rescaled to a percentage of its 1000-point maximum, the 2000-point MME-P score, and answer accuracy for SEED-Bench-2-Plus (abbreviated SEED-2 in the tables, following GenLIP) and ScienceQA. For captioning on the NoCaps and TextCaps validation sets we use the prompt "Provide a one-sentence caption for the provided image." with 5-beam search and up to 64 new tokens, and report CIDEr ($\times100$) computed by pycocoevalcap.

C Additional Experiments

C.1 Training Cost

Table 7 summarizes the training budgets, GPU-hours, and peak memory of the three recipes. All costs are measured on GenLIP-L using NVIDIA L20 GPUs, with batch sizes following each method's training recipe. DIP and NeCo follow their published COCO recipes, and GLASS uses one epoch of Recap-DataComp-12M. GLASS requires 65.5 GPU-hours, compared with NeCo's 156.1, with the masked student pass, full-image EMA teacher, and separate layout-recovery pass all included in its training cost. GLASS generates its recovery targets online and requires no offline pseudo-label preparation. The reported GPU-hours exclude offline preprocessing for all methods, including DIP's diffusion-based pseudo-mask generation. Token counts follow the conventions specified in the table caption.

C.2 Training-Free Retrieval Segmentation

The dense results in the main text rely on linear probes. Here we add the training-free retrieval evaluation that DIP (Sirko-Galouchenko et al., 2025) targets, namely non-parametric dense nearest-neighbor segmentation, which evaluates label transfer through frozen patch-feature neighborhoods without a learned task-specific head. We follow the DIP protocol.

All methods are evaluated at an input resolution of 448×448, yielding a 28×28 patch grid for patch size 16 and a 32×32 grid for patch size 14. A memory bank of patch features is built from the

Table 7: **Training cost comparison.** GPU-hours is effective training time times the number of GPUs and excludes offline preparation; DIP's GPU-hours in particular omit the corpus-scale diffusion pass that builds its pseudo-masks. †Total Tok for GLASS counts both image and text (caption) tokens because GLASS retains the LM objective.

Method	Images (M)	Enc. Tok/Img	Total Tok (B)	Mem. (GB)	GPU-hours
DIP	4.9	196	1.0	6.4	12.4
NeCo	5.9	928	5.5	22.0	156.1
GLASS	12.0	490	6.6†	21.8	65.5

Table 8: **Training-free retrieval segmentation under the DIP protocol.** Dense nearest-neighbor retrieval mIoU on PASCAL-VOC and ADE20K at 448×448 input resolution and support-set fractions $1/1$, $1/8$, $1/64$, and $1/128$. Rows highlighted with denote our method.

Method	Arch	PASCAL-VOC ↑				ADE20K ↑			
		$1/1$	$1/8$	$1/64$	$1/128$	$1/1$	$1/8$	$1/64$	$1/128$
SigLIP 2	g/16	63.6	55.8	42.8	36.3	33.7	26.9	20.0	17.4
GenLIP	L/16	64.7	57.9	45.4	38.6	35.1	28.8	21.8	19.1
+ DIP	L/16	64.9	58.2	45.8	38.9	35.2	28.9	21.9	19.2
+ NeCo	L/16	76.4	**74.0**	**65.0**	**60.7**	42.3	35.6	**28.5**	**26.0**
+ GLASS	L/16	**76.7**	73.0	64.8	60.5	**42.6**	**35.7**	28.2	25.4
OpenVision 2	L/14	59.1	51.4	41.4	37.0	32.3	25.9	20.0	17.7
+ DIP	L/14	69.4	66.8	61.1	57.3	37.1	31.1	24.8	22.2
+ NeCo	L/14	75.9	72.8	**68.6**	**63.4**	39.6	33.2	25.6	22.7
+ GLASS	L/14	**77.6**	**73.9**	65.6	57.4	**41.3**	**33.9**	**26.3**	**23.1**

support split, trainaug for PASCAL-VOC (Everingham et al., 2010) and train for ADE20K (Zhou et al., 2017), where each bank patch carries a soft label derived from its pixel class histogram. For every query patch on the validation split, we retrieve its 30 nearest bank patches by Euclidean distance without ℓ_2-normalizing the query or bank features. Only after neighbor selection are the query and retrieved features normalized within the cosine-similarity scoring step. We aggregate the neighbors' soft labels with $\mathrm{softmax}(\cos/\tau)$ weights at τ=0.07, bilinearly upsample the resulting label map to the input resolution, and score it with pixel-level mIoU under the standard ignore masking. Support features are extracted from two augmented views per image, the bank is capped at $10{,}240{,}000$ patches with rarity-weighted subsampling, and each $1/N$ subset is drawn by greedy class-balanced image selection and averaged over five seeds. Table 8 reports PASCAL-VOC and ADE20K mIoU at support fractions $1/1$, $1/8$, $1/64$, and $1/128$ for GenLIP and OpenVision 2 under the same post-training recipes as Table 1, testing whether the dense gains measured by linear probes carry over to the retrieval regime that DIP is designed for.

Retrieval results. Table 8 shows that GLASS improves over pretraining on both datasets and both encoders at every evaluated support fraction. With full support, GLASS achieves the highest mIoU within each generative-encoder block on both PASCAL-VOC and ADE20K. With reduced support, the comparison with NeCo varies by dataset: GLASS retains higher scores on OpenVision 2 ADE20K, while NeCo has higher PASCAL-VOC scores at $1/64$ and $1/128$. DIP also improves over pretraining, with larger gains on OpenVision 2 than on GenLIP. Overall, the improvements over pretraining extend from linear probing to nearest-neighbor label transfer across a range of support-set sizes.

C.3 DIP WITH FULL FINE-TUNING

The DIP baseline updates only the last three backbone blocks following its original recipe (Sirko-Galouchenko et al., 2025). Table 9 compares it with full-backbone fine-tuning on GenLIP-L/16, keeping other DIP settings unchanged and using the protocols in Appendix B. Full fine-tuning improves all four dense metrics: the ADE20K gap to GLASS narrows from 4.0 to 0.7 points, and the NYUv2 RMSE gap from 0.068 to 0.018. However, ImageNet accuracy falls from 82.8 to 81.9 and all four multimodal scores decrease. GLASS scores better than both DIP variants on all nine

Table 9: **DIP with full fine-tuning on GenLIP-L/16.** DIP updates only the last three backbone blocks following the original recipe (Sirko-Galouchenko et al., 2025), whereas DIP† updates the full backbone with all other DIP settings unchanged. GLASS is included for comparison across dense prediction, global recognition, and multimodal understanding.

Recipe	Dense Task				Global Task	Multimodal understanding			
	ADE20K ↑	Cityscapes ↑	NYUv2 ↓	KITTI ↓	IN Classif. ↑	ChartQA ↑	OCR-B ↑	MME-P ↑	SQA ↑
Pretrained	41.0	61.6	0.507	3.554	82.8	41.2	**51.1**	1258	**76.1**
DIP	43.2	62.0	0.506	3.514	82.8	39.3	49.8	1248	73.4
DIP†	46.5	67.8	0.456	3.416	81.9	36.8	46.6	1231	68.0
GLASS	**47.2**	**70.1**	**0.438**	**3.277**	**82.9**	**42.4**	51.0	**1264**	**76.1**

reported metrics. Updating more blocks therefore narrows the dense gap in this configuration, but does not reproduce GLASS's observed balance with downstream multimodal performance.

C.4 Image–Caption Correspondence Diagnostic

Figure 1 shows that anti-semantic masking increases cross-image feature similarity and degrades dense prediction without a comparable change in caption loss. We examine sensitivity to image–caption correspondence more directly by measuring how replacing the correct image with mismatched images changes teacher-forced caption negative log-likelihood (NLL).

Protocol. We compare the final GenLIP-L/16 student checkpoints trained with semantic, random, and anti-semantic masking under the same LM, MR, and LR configuration, using the official `genlip-l-navit` checkpoint as the pretrained reference. All checkpoints are evaluated on the same 1,000 COCO validation image–caption pairs. Each caption is paired with its correct image and $K{=}5$ distinct mismatched images sampled uniformly from the other selected images. Image selection, caption selection, and mismatched-image sampling use seed 0, with the resulting sample and pairing lists shared across checkpoints. We evaluate each checkpoint through its original caption-prediction pathway with all parameters fixed. Each model receives complete images at 224×224, using the same preprocessing, tokenizer, caption prefixes, and target-token mask across checkpoints and image conditions.

NLL and image sensitivity. For a caption y paired with image x, we define

$$\ell_\theta(x, y) = -\frac{1}{S} \sum_{t=1}^{S} \log p_\theta(y_t \mid y_{<t}, x), \tag{10}$$

where S counts caption target tokens including EOS, but excludes padding, BOS, and non-caption prompt tokens. The mean S in the pretrained evaluation is 13.66. NLL is measured in nats per token. Holding the caption and its teacher-forced prefixes fixed, the image-sensitivity gap is

$$\Delta_{\text{img}}(x, y; \theta) = \frac{1}{K} \sum_{k=1}^{K} \ell_\theta(x'_k, y) - \ell_\theta(x, y), \tag{11}$$

where x'_k is the kth mismatched image. We report the mean correct-image NLL, the mean mismatched-image NLL, and the mean gap across the 1,000 image–caption pairs. The 95% confidence interval for each mean gap uses 10,000 bootstrap resamples of images, taking the 2.5th and 97.5th percentiles. All five mismatches remain grouped with their caption, and the same resampled image indices are used across checkpoints. These intervals summarize image-level uncertainty for each fixed checkpoint.

We additionally report Retrieval@1, the percentage of captions for which the correct image has the lowest NLL among the six candidates (one correct and five mismatched images). Uniform random selection has an expected accuracy of $1/6 \approx 16.7\%$. This ranking-based measure complements the mean NLL gap within the same candidate sets.

Results. Table 10 shows that anti-semantic masking changes correct-image NLL from 3.956 to 4.014, an increase of 0.058 relative to pretraining, while mismatched-image NLL falls from 5.566

Table 10: **Image–caption correspondence diagnostic on GenLIP-L/16.** All checkpoints use the same image–caption pairs and mismatched candidates. NLL and its gap are in nats per token, with image-bootstrap 95% confidence intervals for the mean gap. Retrieval@1 reports correct-image selection among six candidates as a percentage.

Checkpoint	Correct NLL	Mismatched NLL	Δ_{img}	95% CI of gap	Retrieval@1 (%)
Pretrained	3.956	5.566	+1.609	[1.573, 1.646]	97.6
Semantic (GLASS)	3.860	5.537	+1.677	[1.636, 1.718]	97.5
Random	3.932	5.351	+1.419	[1.380, 1.458]	95.2
Anti-semantic	4.014	4.652	+0.638	[0.599, 0.676]	75.7

to 4.652. The mean image-sensitivity gap consequently shrinks from 1.609 to 0.638, a reduction of approximately 60.3%, and Retrieval@1 falls from 97.6% to 75.7%. Thus, the model penalizes mismatched images less and selects the correct image less often, despite the comparatively small change in correct-image NLL. Semantic masking yields a gap of 1.677 and Retrieval@1 of 97.5%, close to the pretrained values; random masking gives intermediate values of 1.419 and 95.2%.

Interpretation. These results show that image–caption correspondence sensitivity can decline substantially while correct-pair caption NLL changes relatively little. Anti-semantic masking retains image-dependent information, as indicated by its positive gap and above-chance Retrieval@1, but responds less strongly to whether the paired image is correct. Together with the increased cross-image feature similarity and dense degradation in Figure 1 and Table 4, this pattern is consistent with increased reliance on text priors. Semantic masking retains near-pretrained correspondence sensitivity alongside improved dense performance, supporting the choice of caption-relevant visible context in GLASS.

D Future Work

Future work could extend caption-guided recovery to video by incorporating temporal correspondences alongside spatial layout. Another direction is to develop more direct measures of visual reliance during post-training, using them to guide masking when captions provide incomplete or noisy descriptions of the visual content.

E Qualitative Results

Figure 6 compares the visible sets selected by random, anti-semantic, and semantic masking on the same image–caption pairs. Figures 4 and 5 compare segmentation and depth predictions across the five dense benchmarks. The examples complement the quantitative results by showing differences in region consistency, object boundaries, and depth structure. All predictions use frozen encoders with task-specific linear probes under the evaluation protocols in Appendix B.

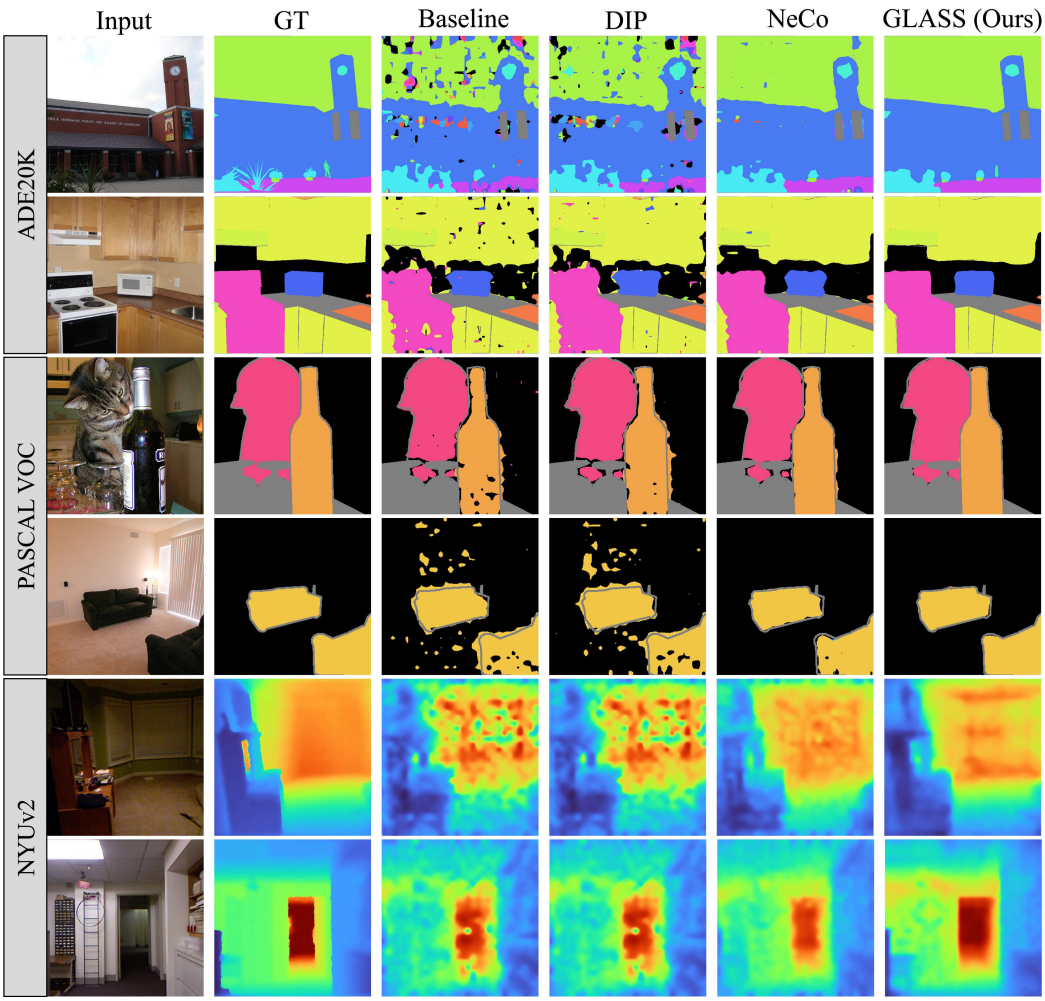

Figure 4: **Qualitative results on general scenes and indoor depth.** Columns show the input image, ground truth (GT), the pretrained encoder (*Baseline*), and the DIP, NeCo, and GLASS post-trained variants. Segmentation ignore regions are displayed in gray using the same ground-truth mask for all methods. Depth maps are shown in pseudocolor.

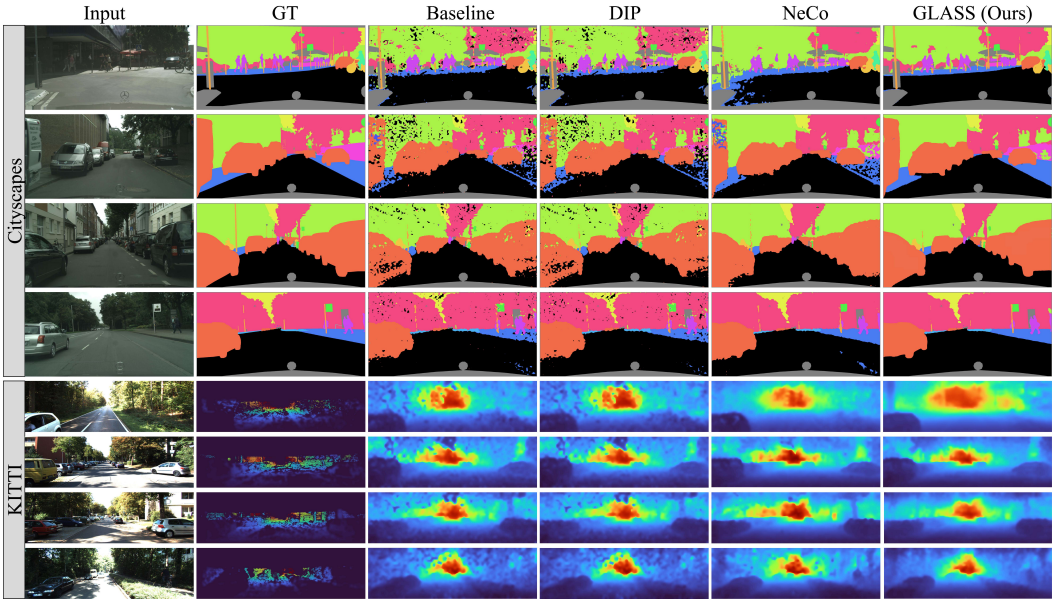

Figure 5: **Qualitative results on driving scenes.** Gray regions in the segmentation maps denote ignored pixels, including the ego vehicle, and are masked identically for all methods only for display. KITTI ground truth is sparse; its dark unmeasured regions do not indicate near-zero depth. Model predictions are displayed as dense depth maps.

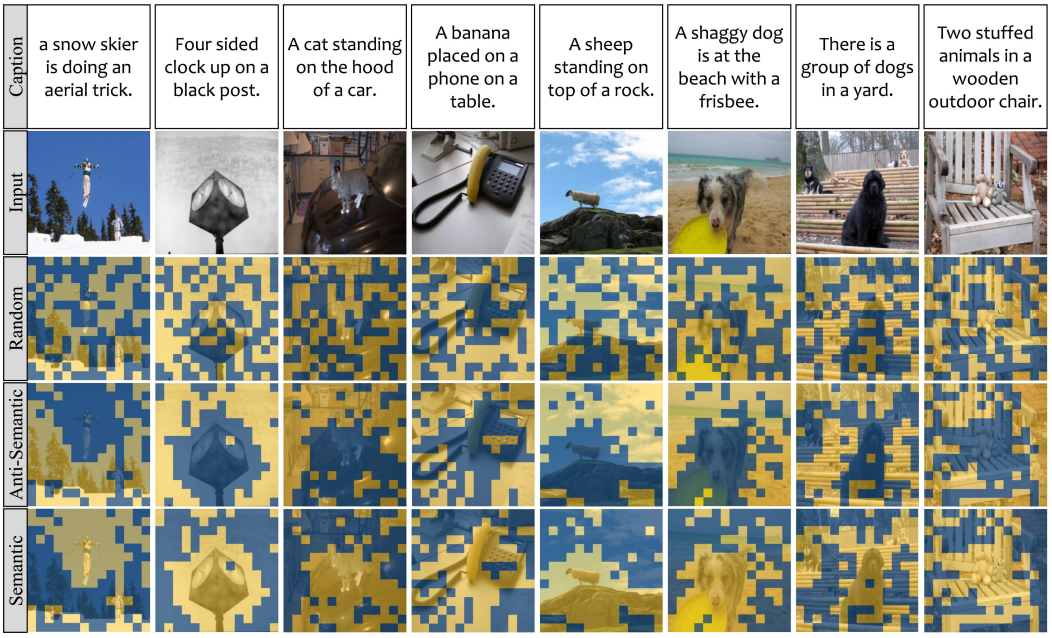

Figure 6: **Comparison of patch selection rules.** Each column shows an image–caption pair, followed by random, anti-semantic, and semantic masking. Yellow overlays denote visible patches, and blue overlays denote masked patches. All three rules keep 50% of the patches visible. During training, masked patch embeddings are replaced with a learnable mask token.