

Beyond Caption Loss: Generative Distillation for Modality Alignment in Vision-Language Models

Nam Kha Nguyen¹ Wei Emma Zhang¹ Xiang Dai²

Mong Yuan Sim^{1,2} Weitong Chen¹

¹Adelaide University

²CSIRO

{namkha.nguyen;wei.e.zhang}@adelaide.edu.au

Abstract

Vision-language models align frozen vision encoders to language models using only a next-token prediction loss, an asymmetric signal that discards rich visual signals not captured in the captions. We propose GENDA (Generative Distillation Alignment), a novel training framework that augments the captioning loss with a generative distillation objective from a frozen text-to-image diffusion model. GENDA concatenates selected visual tokens with the encoded caption to form a joint conditioning signal, encouraging these tokens to encode visual information not already captured by the caption. Across five vision encoders and four language models, GENDA gains 3–9 CIDEr points over LLaVA-1.5 on five captioning benchmarks. These gains also transfer to downstream VLM benchmarks while preserving the encoders’ discriminative performance. Visual preservation and modality-gap analyses suggest that visual information is retained deeper within the LLM, while visual features become more closely aligned with the language representation space.¹

1 Introduction

Most modern Vision-Language Models (VLMs) follow a standard paradigm that a pretrained vision encoder is joined to a pretrained Large Language Model (LLM) through a small projection module (e.g., Q-Former (Li et al., 2023b) or MLP (Liu et al., 2024a)). Training proceeds in two steps: (i) *feature alignment*, in which visual features are aligned to the LLM’s input space, and (ii) *visual instruction tuning*, in which curated multimodal instruction data are used to train the model for vision-language understanding and instruction following. In both stages, the vision encoder, usually a CLIP or SigLIP ViT (Radford et al., 2021; Zhai et al., 2023), stays frozen, under the assumption

¹Code is available at github.com/noonamkha/GenDA

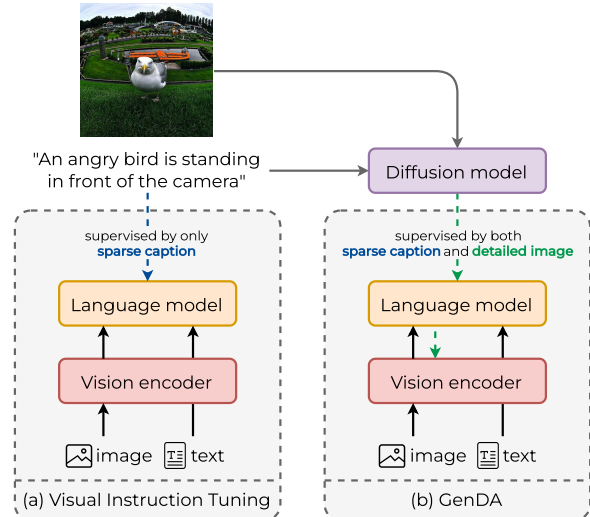


Figure 1: **Standard alignment vs. GENDA.** (a) Conventional visual instruction tuning only supervises through the next-token loss of a sparse caption. (b) GENDA additionally routes selected visual tokens into a diffusion model conditioned on the same caption, so the encoder is supervised by both the caption and the source image.

that a projection module is sufficient to bridge the visual and textual representation spaces. LLaVA (Liu et al., 2023a, 2024a), MiniGPT-4 (Zhu et al., 2024), InstructBLIP (Dai et al., 2023a), Qwen-VL (Bai et al., 2023), and many other VLMs follow this paradigm.

The recipe works, but it has a built-in problem: the two modalities are usually not symmetric. For example, in image captioning, images carry far more information than captions describe (Schrodi et al., 2025). Pictures may contain textures, fine spatial relations, and small attribute differences, whereas captions do not explicitly describe everything in detail. When the only training signal comes from predicting the next text token, the gradient pushes visual features toward whatever maps cleanly onto words, which means anything that does not appear in the caption is pushed out. The

vision encoder ends up trying to fit a narrowed text space, but the alignment loss is not strong enough to close the gap (Liang et al., 2022).

We argue that this gap should be closed by complementing the language signal with a second source of supervision that is sensitive to the visual details that captions leave out. Pretrained text-to-image diffusion models such as (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024) can offer such a signal. They are trained to generate pixels from text, and their text-conditioning space has two properties at once: it is aligned with natural language (the input is text), and it is rich enough to specify an image down to fine details (the output is pixels). The diffusion model’s score of an image conditioned on a caption gives a dense, text-aware training signal that a vision encoder can be pushed toward. This makes the diffusion conditioning space a useful meeting point between vision and language

Building on this idea, we propose GENDA, an improved feature alignment method that adds a diffusion-based generative loss to the standard captioning loss. Figure 1 contrasts our approach with the conventional pipeline. During the first stage of VLM training, the vision encoder’s output is sent down through two paths. The first path is the usual path through the projector and into the LLM, supervised by next-token prediction. The second path projects a small, informative subset of the visual tokens into the diffusion conditioning space, joins them with the encoded caption, and employs a frozen diffusion model to generate the source image from this combined signal. The resulting noise-prediction loss flows back through the projection and into the vision encoder. Because the caption is already provided to the diffusion model, the visual tokens only earn gradients when they carry information the caption misses, pushing the encoder to preserve precisely the perceptual detail that text supervision throws away.

In summary, our contributions are as follows:

- We propose GENDA, which reframes VLM feature alignment as a two-signal problem, combining a language signal from the LLM with a complementary visual signal from a frozen diffusion model to address the asymmetry of caption-only supervision.
- We demonstrate consistent performance gain across five image captioning benchmarks,

with the combination of five vision encoders, and four LLMs while preserving zero-shot retrieval and classification performance.

- We show via layer-wise Centered Kernel Nearest-Neighbor Alignment (CKNNA) and modality-gap visualization that generative feedback preserves visual information deeper into the LLM and pulls visual tokens closer to the text-token space.

2 Related Work

Vision-Language Models Training.

The LLaVA family (Liu et al., 2023a, 2024a) showed that a plain two-layer MLP trained in the two stages described above is enough to get a strong multimodal assistant when paired with a capable LLM. In this paradigm, the vision encoder is frozen during alignment, with CLIP, SigLIP, or other variants such as OpenCLIP (Cherti et al., 2023) used as fixed feature extractors. Tong et al. (2024) show that VLM errors often come from the encoder’s inability to tell apart fine-grained visual details. Current approaches to addressing this issue generally fall into two categories: combining features from multiple frozen encoders (Kar et al., 2024; Shi et al., 2025), or pretraining encoders with high-quality image-text data (Xu et al., 2024; Cherti et al., 2023). These methods give stronger frozen backbones but they do not change what happens during multimodal alignment. There, the encoder is still a fixed feature source and any remaining mismatch with the LLM has to be absorbed by a projector with limited capacity.

Diffusion Models as Representation Learner.

Following the empirical success of large text-to-image diffusion models, several lines of work have asked whether the representations these models learn for generation are useful for other tasks. DIFT (Tang et al., 2023) and Diffusion Hyperfeatures (Luo et al., 2023) extract intermediate UNet activations and show they encode dense semantic correspondences. In the multimodal direction, DreamLLM (Dong et al., 2024), Emu (Dai et al., 2023b), and SEED-LLaMA (Ge et al., 2024) attach diffusion decoders to LLMs so that the composed model can generate images.

Closest to our work are methods that directly distill from diffusion models. DreamTeacher (Li et al., 2023a) distills diffusion features into discriminative backbones, while DIVA (Wang et al.,

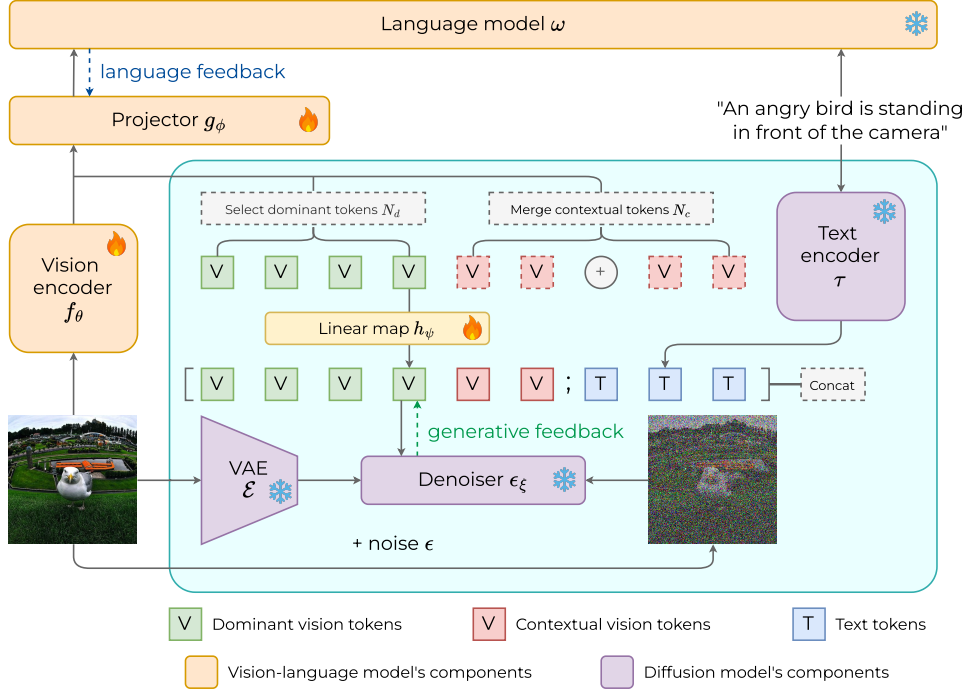


Figure 2: **Overall architecture of GENDA.** Visual tokens flow along two paths: the standard projector-LLM path trained with the captioning loss (*language feedback*), and a parallel path in which dominant tokens and merged contextual tokens are linearly mapped into the diffusion conditioning space, concatenated with the encoded caption, and passed to a frozen denoiser to reconstruct the source latent (*generative feedback*).

2025) refines CLIP via a frozen diffusion model’s reconstruction objective. These prior works use diffusion models to improve standalone vision backbones without incorporating textual information. GENDA differs in that the diffusion loss is text-conditional (computed under the ground-truth caption) and is applied to an encoder inside a larger VLM trained jointly with a captioning loss, making it part of multimodal alignment rather than a separate pixel-reconstruction stage. As far as we know, no prior method uses a frozen text-to-image diffusion model as a text-conditional regularizer during VLM feature-alignment training stage.

3 Methodology

Our proposed method augments the standard feature-alignment stage of VLM pretraining with a generative distillation objective derived from a frozen diffusion model. The overall architecture is depicted in Figure 2. We first establish notations, then describe token selection, the generative distillation objective, and the joint training procedure.

3.1 Notations and Preliminaries

Let $f_\theta(\cdot)$ denote the vision encoder (e.g., a CLIP-pretrained ViT), $g_\phi(\cdot)$ the connector – typically a two-layer MLP – that projects visual features

into the LLM embedding space, and $h_\psi(\cdot)$ a lightweight linear projector that maps visual features into the conditioning space of a frozen diffusion model. We further write $\tau(\cdot)$ for the diffusion model’s frozen text encoder, $\epsilon_\xi(\cdot)$ for its frozen denoising network, and $\mathcal{E}(\cdot)$ for its VAE encoder. The downstream LLM, with parameters ω , also remains frozen during alignment pretraining. We provide a brief overview of the diffusion model formulation used throughout this work in Appendix A.

Given an image-caption pair (x, c) , the vision encoder produces L visual tokens $v = f_\theta(x) \in \mathbb{R}^{L \times D_v}$ where D_v denotes the dimension of visual token embedding. In the standard LLaVA-style pipeline, $g_\phi(v)$ is prepended to the tokenized caption and supervised under the autoregressive language model (LM) loss:

$$\mathcal{L}_{\text{LM}} = - \sum_{i=1}^{|c|} \log p_\omega(c_i | g_\phi(v), c_{<i}), \quad (1)$$

whose gradients flow through g_ϕ into f_θ , providing the encoder’s only learning signal during alignment. Our aim is to enrich this signal by introducing generative feedback from diffusion models.

3.2 Token Selection

Of the L visual tokens, many correspond to background patches that contribute little to reconstruction, yet naively discarding them removes contextual cues the diffusion model benefits from. Inspired by VisionZip (Yang et al., 2025), we compress v into $N \ll L$ tokens via *dominant token selection* and *contextual token merging*.

Dominant token selection. Let a_j denote the attention that token v_j receives from the global token in the final transformer block of f_θ (the [CLS] token for CLIP, the attention-pooling token for SigLIP). We retain the N_d dominant tokens with the highest scores:

$$\begin{aligned} \mathcal{D} &= \text{TopK}(\{a_j\}_{j=1}^L, N_d), \\ v^d &= \{v_j : j \in \mathcal{D}\} \in \mathbb{R}^{N_d \times D_v}. \end{aligned} \quad (2)$$

These patches tend to concentrate object-level and salient-attribute information.

Contextual token merging. To preserve background context, we merge the remaining tokens rather than discard them. Let $\mathcal{R} = \{1, \dots, L\} \setminus \mathcal{D}$. We partition $\mathcal{R}$ into N_c groups $\{\mathcal{G}_k\}_{k=1}^{N_c}$ by clustering key embeddings under cosine similarity, and average within each group:

$$v_k^c = \frac{1}{|\mathcal{G}_k|} \sum_{j \in \mathcal{G}_k} v_j, \quad k = 1, \dots, N_c. \quad (3)$$

The diffusion projector receives

$$\tilde{v} = [v^d; v^c] \in \mathbb{R}^{N \times D_v}, \quad N = N_d + N_c \ll L. \quad (4)$$

The procedure reduces token count by roughly an order of magnitude with negligible overhead. Alternative selection strategies are examined in Section 4.4.

3.3 Generative Distillation Objective

The selected tokens are projected into the diffusion conditioning space via $h_\psi(\tilde{v}) \in \mathbb{R}^{N \times D_\tau}$, where D_τ is the output dimensionality of the diffusion model’s frozen text encoder τ . We construct the joint conditioning signal by concatenating the projected visual tokens with the encoded caption,

$$\mathbf{c}_{\text{joint}} = [h_\psi(\tilde{v}); \tau(c)] \in \mathbb{R}^{(N+|c|) \times D_\tau},$$

and supply $\mathbf{c}_{\text{joint}}$ to the frozen denoising network in place of a purely textual prompt.

The input image is encoded into a latent $z_0 = \mathcal{E}(x)$, perturbed at a uniformly sampled timestep $t \sim \mathcal{U}\{1, \dots, T\}$ with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, yielding $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ under the standard variance-preserving schedule. The generative distillation loss is the diffusion model’s native noise-prediction objective evaluated under the joint conditioning:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{z_0, \epsilon, t} [\|\epsilon - \epsilon_\xi(z_t, t, [h_\psi(\tilde{v}); \tau(c)])\|_2^2]$$

Unlike DIVA (Wang et al., 2025) which conditions the frozen denoiser on $h_\psi(\tilde{v})$ alone, we concatenate $h_\psi(\tilde{v})$ with the encoded caption $\tau(c)$. The two methods optimize for different goals: DIVA refines a standalone CLIP encoder, whereas our encoder is always consumed jointly with text. Without $\tau(c)$ as a conditioner, the encoder would be pressured to re-encode information that the caption already provides, wasting capacity on redundant content. Concatenation makes the visual tokens responsible for precisely the residual information that the caption fails to convey: the texture, layout, and attribute detail that constitute the modality gap. The diffusion model thus functions as a differentiable critic that scores how completely the visual tokens, together with the caption, specify the source image. We examine the importance of $\tau(c)$ in Section 4.4.

3.4 Joint Training and Composite Objective

The vision encoder f_θ , the LLM connector g_ϕ , and the diffusion projector h_ψ are trained jointly. The diffusion backbone ϵ_ξ , its text encoder τ , its VAE encoder $\mathcal{E}$, and the LLM all remain frozen throughout alignment. The composite objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{gen}},$$

where $\lambda \in \mathbb{R}_{\geq 0}$ controls the trade-off between linguistic alignment and generative visual fidelity.

At inference time, the diffusion pathway—including h_ψ , τ , $\mathcal{E}$, and ϵ_ξ —is discarded. The deployed model is architecturally identical to a standard LLaVA-style VLM, so the cost of GENDA is borne entirely at training time and the resulting checkpoint is a drop-in replacement for conventionally pretrained alignment weights. The second stage, visual instruction tuning, proceeds without modification.

Table 1: **Image captioning results (CIDEr) on five benchmarks.** Comparison of LLaVA-1.5 and GENDA across five vision encoders (CLIP-L/14, InternViT-300M (Gao et al., 2024), OpenCLIP-L/14 (Cherti et al., 2023), EVA-02-CLIP-L/14 (Sun et al., 2023), SigLIP-SO-400M (Zhai et al., 2023)) and four language models (Vicuna-7B/13B (Zheng et al., 2023), Llama3-8B (Grattafiori et al., 2024), Qwen2-7B (Yang et al., 2024)). GENDA consistently gains 3-9 CIDEr points on average, with the largest improvements on VizWiz-Caps and stronger vision backbones.

Method	Vision encoder	Language model	COCO	Flickr	nocaps	TextCaps	VizWiz
Baseline							
LLaVA-1.5	CLIP-L/14	Vicuna-7B	91.2	58.4	87.2	79.8	66.1
GENDA	CLIP-L/14	Vicuna-7B	93.8	61.7	91.4	79.2	70.4
LLaVA-1.5	CLIP-L/14	Vicuna-13B	104.5	63.9	92.3	81.6	71.4
GENDA	CLIP-L/14	Vicuna-13B	108.2	69.4	98.4	80.2	75.9
Expand to other LLMs							
LLaVA-1.5	CLIP-L/14	Llama3-8B	106.1	64.6	98.6	82.5	77.7
GENDA	CLIP-L/14	Llama3-8B	109.7	67.9	101.4	82.9	81.2
LLaVA-1.5	CLIP-L/14	Qwen2-7B	110.3	68.2	103.8	84.1	83.5
GENDA	CLIP-L/14	Qwen2-7B	112.5	72.8	109.7	83.6	87.0
Expand to more robust vision encoders							
LLaVA-1.5	InternViT-300M	Vicuna-7B	93.6	62.5	87.3	82.2	71.8
GENDA	InternViT-300M	Vicuna-7B	98.3	67.8	92.1	81.9	78.3
LLaVA-1.5	OpenCLIP-L/14	Vicuna-7B	93.8	63.1	91.7	84.4	72.3
GENDA	OpenCLIP-L/14	Vicuna-7B	101.9	70.3	98.1	84.7	79.2
LLaVA-1.5	EVA-02-CLIP-L/14	Vicuna-7B	96.2	64.4	93.3	85.2	74.9
GENDA	EVA-02-CLIP-L/14	Vicuna-7B	105.7	73.8	100.9	88.4	81.5
LLaVA-1.5	SigLIP-SO-400M	Vicuna-7B	99.6	66.3	94.7	87.9	76.3
GENDA	SigLIP-SO-400M	Vicuna-7B	110.9	76.4	102.8	91.4	85.5

4 Experiments

We follow the two-stage LLaVA-1.5 protocol (Liu et al., 2024a) and adopt its hyperparameters unless otherwise noted. Stage 1 (feature alignment) is performed on LLaVA-558K and Stage 2 (visual instruction tuning) on LLaVA-665K. Our default configuration uses CLIP-L/14 (Radford et al., 2021) as the vision encoder, Vicuna-7B (Zheng et al., 2023) as the language model, Stable Diffusion 1.5 (Rombach et al., 2022) as the frozen diffusion model. For each configuration, we select λ at the beginning of training using a LR-finder-style sweep (Smith, 2017). For the token selection module, we select 54 dominant tokens along with 10 contextual tokens. Further details about training implementation is provided in Appendix B and C.

4.1 Main result

We first evaluate the immediate effect of generative distillation on the encoder-LLM alignment by running image captioning directly after Stage 1 on COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), nocaps (Agrawal et al., 2019), TextCaps (Sidorov et al., 2020) and VizWiz-Captions (Gurari et al., 2020). Results are reported in Table 1. GENDA yields consistent CIDEr improvements over LLaVA-1.5 on most benchmarks across every backbone configuration we tested, with average gains of 3-9 points. Two patterns are worth high-

lighting. First, GENDA gains the most on VizWiz-Caps (e.g., +6.5 with CLIP-L/Vicuna-7B and +9.2 with SigLIP-SO/Vicuna-7B), a dataset featuring images with unusual viewpoints captured by blind users. This is where the text-conditional diffusion signal is expected to matter: the encoder must retain perceptual detail that captions rarely express. Second, the gains grow with the strength of the vision encoder, from +2.6 COCO CIDEr on CLIP-L/14 to +11.3 on SigLIP-SO-400M, indicating that the generative signal does not only compensate for a weak backbone but also improve representational capacity that the standard captioning loss leaves unused. The improvements also transfer across language models. Pairing CLIP-L/14 with Llama3-8B and Qwen2-7B yields gains of comparable magnitude to those obtained with Vicuna (+2.2 to +3.6 CIDEr on COCO), indicating that the benefits are encoder-side and not tied to a particular LLM.

GENDA underperforms the baseline on some configuration of TextCaps benchmark (e.g., 79.2 vs. 79.8 with CLIP-L/Vicuna-7B). TextCaps demands accurate transcription of scene text, and SD-1.5 is known to render text poorly. Liu et al. (2023b) shows that SD-1.5 only generates text regions as generic texture. Its loss therefore rewards tokens that encode texture and penalizes tokens that encode the actual characters, so the encoder

Table 2: **Zero-shot retrieval and image classification with the GENDA-trained encoders.** Text/image retrieval on COCO and Flickr30K, and image classification on ImageNet-1k (IN-1k), CIFAR-10/100 (C-10/C-100), and Caltech-101 (Cal-101). Across all five backbones, GENDA matches or slightly exceeds the original frozen weights, showing that generative distillation does not cause catastrophic drift of the encoder’s discriminative behavior.

Method	Text retrieval		Image retrieval		Image classification			
	COCO	Flickr	COCO	Flickr	IN-1k	C-10	C-100	Cal-101
CLIP-L	56.3	85.3	36.5	65.0	75.5	95.6	77.9	92.6
Ours	56.4	85.8	36.5	65.4	75.9	95.6	76.2	92.8
InternViT-300M	58.7	86.4	38.1	66.8	76.2	96.1	79.4	93.1
Ours	58.8	86.8	38.6	67.1	76.2	96.3	79.5	92.9
OpenCLIP	59.4	86.8	39.0	67.3	75.3	96.6	83.4	92.9
Ours	59.8	87.2	39.3	67.5	75.7	97.0	83.6	93.2
EVA-CLIP	64.2	89.3	47.9	78.7	79.8	99.1	91.1	94.5
Ours	64.1	89.5	48.2	78.9	79.8	99.2	91.4	94.8
SigLIP	70.2	91.5	52.1	81.6	80.5	98.3	88.6	95.2
Ours	70.5	91.8	52.4	81.4	80.8	98.4	88.9	95.3

Table 3: **Downstream VLM benchmarks after Stage-2 visual instruction tuning.** GENDA improves over LLaVA-1.5 on most benchmarks, with the largest gains on fine-grained spatial/attribute reasoning (GQA, MMVP). OCR-heavy benchmarks (AI2D, OCRVQA, TextVQA) show minor degradation.

Method	Vision encoder	Language model	MMBench	MMStar	MMMU	MMVP	AI2D	OCRVQA	GQA	TextVQA
LLaVA-1.5	CLIP-L	Vicuna-7B	64.3	33.1	35.7	24.7	54.8	58.1	62.0	58.2
Ours	CLIP-L	Vicuna-7B	65.1	34.4	35.9	27.1	54.1	56.2	65.7	56.3
LLaVA-1.5	CLIP-L	Vicuna-13B	68.7	34.3	36.4	30.7	59.5	62.0	63.3	61.3
Ours	CLIP-L	Vicuna-13B	69.2	36.8	37.1	33.6	58.7	63.2	66.5	60.7

is trained to discard OCR features, which leads to the negative effect on TextCaps. We verify this hypothesis in Section 4.4 by replacing SD-1.5 with more capable diffusion backbones.

4.2 Downstream tasks

Vision-encoder tasks. A usual concern when injecting a learning signal into a pretrained encoder is that the encoder may drift away from the visual properties that make it useful as a general-purpose feature extractor (Wortsman et al., 2022; Kumar et al., 2022). Table 2 reports zero-shot text/image retrieval on COCO and Flickr and zero-shot classification on ImageNet-1k (Russakovsky et al., 2015), CIFAR-10/100 (Krizhevsky et al., 2009), and Caltech-101 (Fei-Fei et al., 2004). Across all benchmarks, our trained encoders match or slightly exceed the original frozen weights. Generative distillation preserves the encoder’s discriminative behavior, proving that catastrophic drift does not cause the captioning improvements.

Instruction-tuned VLM benchmarks. To verify that Stage-1 results also improve the full LLaVA pipeline, we freeze our Stage-1 encoders and run standard Stage-2 visual instruction tuning on LLaVA-665K. Table 3 reports performance on

eight common VLM benchmarks. GENDA improves on MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024), MMMU (Yue et al., 2024), MMVP (Tong et al., 2024), and GQA (Hudson and Manning, 2019), with the largest gains on GQA and MMVP, benchmarks heavily reliant on fine-grained spatial and attribute reasoning. As in the captioning evaluation, the OCR-heavy benchmarks (AI2D (Hiippala et al., 2021), OCRVQA (Mishra et al., 2019), TextVQA (Singh et al., 2019)) show small degradation, similar to the TextCaps result, suggesting that the limitation is inherited from SD-1.5’s poor text rendering rather than from generative distillation per se. We further extend generative supervision to Stage 2 in Appendix D, and find that it yields only marginal gains over the baseline while underperforming our Stage-1-only design.

4.3 Vision-language alignment

In this section, we evaluate the effectiveness of our method in improving vision-language alignment from two perspectives: *Visual information preservation* and *Modality gap*. Specifically, we evaluate how well visual information flows through the LLM, and how close the visual tokens sit to text tokens in the shared embedding space.

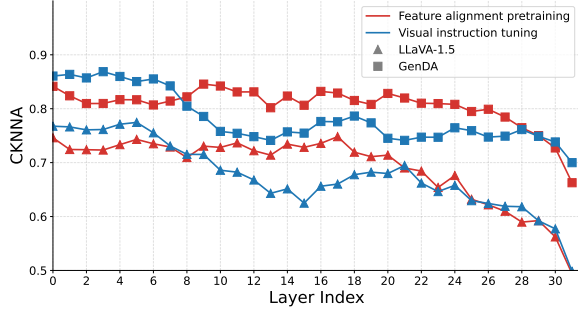


Figure 3: **CKNNA between vision-encoder and LLM visual-token embeddings**, computed after Stage 1 (feature alignment) and Stage 2 (visual instruction tuning). Our method better preserves visual information, with the gap to the baseline widening at deeper layers.

Visual information preservation. Following (Huh et al., 2024), we measure Centered Kernel Nearest-Neighbor Alignment (CKNNA) between the vision encoder’s output and the hidden representation of visual tokens at every layer of the LLM. A higher CKNNA score indicates that the LLM’s per-layer hidden states preserve more of the nearest-neighbor structure of the vision encoder’s output, which we take as a proxy for how much visual information survives at that layer. Figure 3 shows that GENDA maintains higher CKNNA than LLaVA-1.5 after both stages, and the decline across deeper LLM layers is less steep. Our generative objective pressures the encoder to produce tokens whose information is grounded in the image itself, and Figure 3 indicates that this information survives deeper into the LLM’s layers.

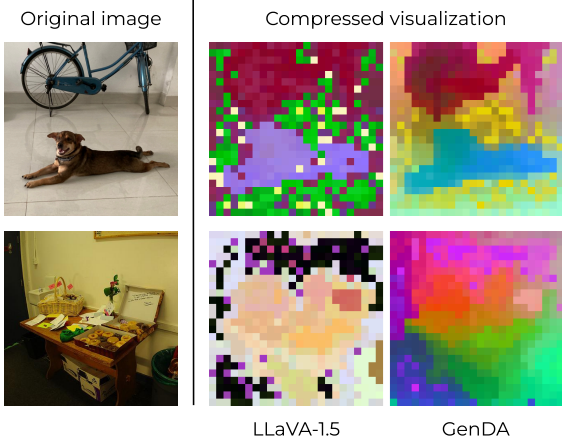


Figure 4: **Per-patch visual-token embeddings**. LLaVA-1.5 (left) versus GENDA (right) on a clean figure-ground image (top) and a cluttered indoor scene (bottom). GENDA yields sharper object-background separation and better preserves spatial regions.

We complement the CKNNA analysis with a qualitative UMAP visualization of per-patch visual-token representations in Figure 4. For images with a clear figure-ground separation, our method can produce more sensible groupings between different objects compared to baseline. For images with a more cluttered indoor scene, our representation can still distinguish major spatial regions (table surface, floor, background wall) that collapse together in the baseline. It shows that the diffusion signal forces the encoder to retain spatial and attribute structure beyond what is needed to describe the object. We quantitatively examine this in Appendix E, where GENDA’s gains increase with scene density and are largest for images containing the most annotated objects.

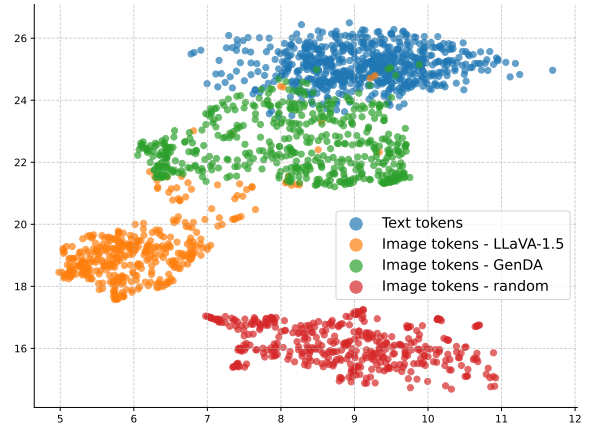


Figure 5: **UMAP of input-token embeddings in the LLM space**. Visual tokens from our model are pulled closer to the text-token cluster than either the LLaVA-1.5 or randomly projected tokens

Modality gap. Liang et al. (2022) show that visual and textual tokens in CLIP-style models occupy disjoint cones in the shared embedding space. We test whether generative distillation narrows this gap. Figure 5 plots a UMAP of four populations: (i) text tokens, (ii) visual tokens from LLaVA-1.5, (iii) visual tokens from our model, and (iv) tokens produced by an untrained, randomly initialized projector. As reported by Liang et al. (2022), all four clusters remain spatially distinct, even random tokens form their own coherent island, indicating that the gap is partly an artifact of representational geometry rather than purely a function of training signal. However, our visual tokens sit closer to the text cluster than either the LLaVA-1.5 tokens or the random tokens, consistent with the Platonic-representation perspective

of Huh et al. (2024) that semantically aligned inputs should converge across modalities. Generative distillation pulls the visual features toward the language space, effectively reducing modality gap.

4.4 Ablation studies

We ablate three design choices: which tokens to expose to the diffusion model, what auxiliary conditioning to include alongside the visual tokens and the choice of diffusion backbone.

Table 4: Ablations on **Tokens selection** and **Auxiliary conditioning**. $\mathcal{S}$ denotes the tokens selection strategy. $\mathcal{T}$ and $\mathcal{G}$ indicates the presence of text tokens and global tokens in auxiliary conditioning respectively. We report CIDEr score of COCO (CC), Flickr30k (FK), nocaps (NC), TextCaps (TC) and VizWiz-Caps (VW)

Method	$\mathcal{S}$	$\mathcal{T}$	$\mathcal{G}$	CC	FK	NC	TC	VW
LLaVA-1.5	-	-	-	91.2	58.4	87.2	79.8	66.1
Ablation on local tokens selection strategy $\mathcal{S}$								
Ours	All	✓	✗	87.6	55.3	83.9	77.2	62.4
Ours	Rand-64	✓	✗	93.4	60.9	91.0	78.6	69.3
Ours	Top-64	✓	✗	93.8	61.7	91.4	79.2	70.4
Ablation on auxiliary conditioning on text $\mathcal{T}$ and global $\mathcal{G}$ tokens								
Ours	Top-64	✗	✗	90.8	57.2	85.6	75.4	64.7
Ours	Top-64	✗	✓	93.1	59.4	89.9	78.2	68.8
Ours	Top-64	✓	✓	92.6	57.5	88.1	77.8	68.3

Token selection. We examine how much and which visual information should be routed through the diffusion pathway. From Table 4, routing all 576 visual tokens through the diffusion pathway *hurts* performance relative to the baseline (e.g., 87.6 vs. 91.2 on COCO). We attribute this to two effects: background patches contribute noisy gradient with little reconstruction signal, and the long conditioning sequence overwhelms the cross-attention of a frozen denoiser tuned for short text prompts. Randomly subsampling 64 tokens already recovers and exceeds the baseline (93.4 on COCO), and selecting the 64 most globally attended tokens yields the strongest result (93.8).

Auxiliary conditioning. We investigate what contextual information the diffusion model should receive alongside the selected visual tokens. Table 4 shows that conditioning the diffusion model on visual tokens alone (no text, no global token) is the weakest variant, underperforming LLaVA-1.5 on every benchmark. Without the caption as a baseline conditioner, the encoder wastes capacity re-expressing information the LLM already gets from text. Consistent with Wang et al. (2025),

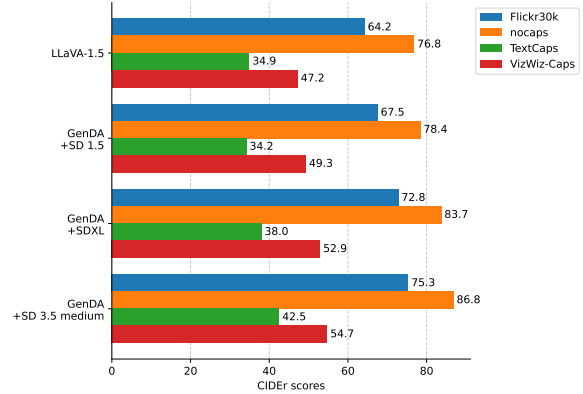


Figure 6: **Captioning performance across diffusion backbones.** Performance scales with diffusion capacity on every benchmark

adding the global token (no text) recovers most of the gap, suggesting it supplies coarse scene context the diffusion model would otherwise infer. Adding text on top of global tokens yields little further gain, indicating the two provide overlapping information; the encoder benefits most when responsible only for the residual neither covers.

Diffusion backbone. We examine whether the quality of the generative model affects the supervision provided by GENDA by replacing SD-1.5 with SDXL (Podell et al., 2024) and SD-3.5-medium (Esser et al., 2024). We restrict this ablation to a smaller-scale setting: Stage-1 pretraining on the COCO Karpathy train split (Karpathy and Fei-Fei, 2015), evaluated on Flickr30K, NoCaps, TextCaps, and VizWiz-Caps. Figure 6 shows that captioning quality improves with the strength of the diffusion model on every benchmark. The most striking effect is on TextCaps that using SD-1.5 show small degradation, both SDXL and SD-3.5-medium produce a clear improvement (+7.5 CIDEr over LLaVA-1.5). This supports our hypothesis from Section 4.1: the TextCaps degradation observed under SD-1.5 is a property of the diffusion model, not of the distillation objective. As text-to-image models continue to improve, GENDA inherits those gains without architectural change. Due to limit hardware, we use SD-1.5 as our main diffusion model for all experiments.

5 Conclusion

We presented GENDA, a feature-alignment framework that augments the standard captioning loss with a generative distillation objective from a frozen text-to-image diffusion model. By concate-

nating selected visual tokens with the encoded caption as joint conditioning, the diffusion model supervises the encoder only on the residual visual signal that text fails to convey. Our experiments show that GENDA improves performance across benchmarks that requires fine-grained understanding, and this result scales with more robust vision encoders. CKNNA and modality-gap analyses suggest that visual information survives deeper into the LLM and that visual tokens move closer to the text-token space.

Limitations

GENDA is applied only during the Stage-1, feature alignment while Stage-2 visual instruction tuning proceeds unchanged. Instruction-tuning data frequently contains short or terse targets (yes/no answers, single-word VQA responses) that describe almost none of the source image, making them weak textual conditioners for the diffusion model and undermining the residual-information interpretation that motivates concatenating the encoded caption. We therefore have not explored whether filtered instruction data could provide useful generative supervision in Stage 2, nor whether unfreezing the vision encoder under our objective during instruction tuning would yield further gains or instead induce forgetting.

Our total loss combines the captioning and generative distillation signals additively, and we did not disentangle their respective contributions to the vision encoder. Both gradients update the encoder and may cooperate, interfere, or dominate one another in ways a single composite objective cannot reveal, with the relative balance unlikely to remain constant across the training trajectory, encoder-LLM pairs, or layers of the encoder. We did not analyze per-layer or per-step gradient norms, nor test gradient-surgery, dynamic re-weighting, or separate optimizers for the two signals. A more principled treatment of this two-signal optimization is left to future work and may help explain the residual variance we observe on OCR-heavy benchmarks.

Acknowledgments

This study was funded by the Australian Research Council Early Career Industry Fellowships (IE230100119) and Linkage Project (LP230200821).

References

- Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [nocaps: novel object captioning at scale](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.
- 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](#). *Preprint*, arXiv:2308.12966.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#) In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. [Reproducible scaling laws for contrastive language-image learning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2818–2829. IEEE.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023a. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, and 7 others. 2023b. [Emu: Enhancing image generation models using photogenic needles in a haystack](#). *Preprint*, arXiv:2309.15807.
- Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. [Dreamllm: Synergistic multimodal comprehension and creation](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. [Scaling rectified flow transformers for high-resolution image synthesis](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR / OpenReview.net.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. [Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories](#). In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2004, Washington, DC, USA, June 27 - July 2, 2004*, page 178. IEEE Computer Society.
- 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. 2024. [Mini-intervl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance](#). *Vis. Intell.*, 2(1):32.
- Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024. [Making llama SEE and draw with SEED tokenizer](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. [Captioning images taken by people who are blind](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 417–434. Springer.
- Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. 2021. [AI2D-RST: a multimodal corpus of 1000 primary school science diagrams](#). *Lang. Resour. Evaluation*, 55(3):661–688.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [De-noising diffusion probabilistic models](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. [Position: The platonic representation hypothesis](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR / OpenReview.net.
- Oguzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. 2024. [BRAVE: broadening the visual encoding of vision-language models](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVI*, volume 15074 of *Lecture Notes in Computer Science*, pages 113–132. Springer.
- Andrej Karpathy and Li Fei-Fei. 2015. [Deep visual-semantic alignments for generating image descriptions](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.
- Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. [Learning multiple layers of features from tiny images](#).
- Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. [Fine-tuning can distort pretrained features and underperform out-of-distribution](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. 2023a. [Dreamteacher: Pretraining image backbones with deep generative models](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16652–16662. IEEE.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. [Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning](#). In *Advances in*

- Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.*
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. 2023b. [Character-aware models improve visual text rendering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16270–16297. Association for Computational Linguistics.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.
- Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023. [Diffusion hyperfeatures: Searching through time and space for semantic correspondence](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. [OCR-VQA: visual question answering by reading text in images](#). In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. [Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. [SDXL: improving latent diffusion models for high-resolution image synthesis](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [Imagenet large scale visual recognition challenge](#). *Int. J. Comput. Vis.*, 115(3):211–252.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. [Photorealistic text-to-image diffusion models with deep language understanding](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. [Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey

- Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. 2025. [Eagle: Exploring the design space for multimodal llms with mixture of encoders](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. [Textcaps: A dataset for image captioning with reading comprehension](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, Lecture Notes in Computer Science, pages 742–758. Springer.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards VQA models that can read](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.
- Leslie N. Smith. 2017. [Cyclical learning rates for training neural networks](#). In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 464–472. IEEE Computer Society.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. [Eva-clip: Improved training techniques for clip at scale](#). *Preprint*, arXiv:2303.15389.
- Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. [Emergent correspondence from image diffusion](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9568–9578. IEEE.
- Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. 2025. [Diffusion feedback helps CLIP see better](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2022. [Robust fine-tuning of zero-shot models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7949–7961. IEEE.
- Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. [Demystifying CLIP data](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. [Visionzip: Longer is better but not necessary in vision language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19792–19802. Computer Vision Foundation / IEEE.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. [From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions](#). *Trans. Assoc. Comput. Linguistics*, 2:67–78.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [Minigpt-4: Enhancing vision-language understanding with advanced large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Diffusion Model Preliminaries

Denoising diffusion probabilistic models (DDPM) learn to reverse a gradual noising process (Ho et al., 2020). The overall forward and reverse diffusion processes are illustrated in Figure 7.

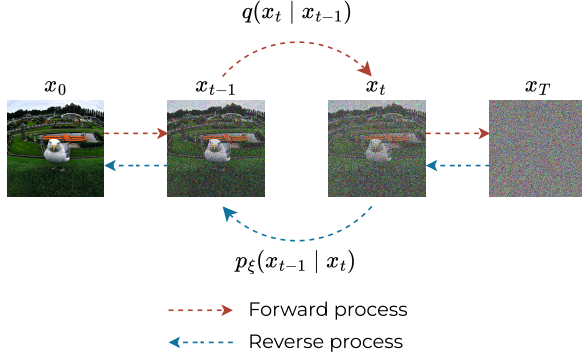


Figure 7: Overview of the diffusion process. The forward process gradually adds Gaussian noise to a clean sample (x_0), while the reverse process iteratively denoises (x_T) to recover the data distribution.

Given a clean data sample x_0 , the forward process adds Gaussian noise over T steps:

$$q(x_t | x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I), \quad (5)$$

where $\alpha_t \in (0, 1)$ determines the noise schedule. Defining $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, a sample at any timestep can be obtained directly as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (6)$$

The reverse process removes noise one step at a time. A denoising network ϵ_ξ receives the noisy sample x_t and timestep t , and predicts the noise added during the forward process. It is commonly trained with the noise-prediction objective

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0, t, \epsilon} \left[\|\epsilon - \epsilon_\xi(x_t, t)\|_2^2 \right] \quad (7)$$

The predicted noise parameterizes the reverse transition $p_\xi(x_{t-1} | x_t)$, whose repeated application transforms Gaussian noise x_T into a sample from the learned data distribution.

Latent diffusion models apply the same diffusion process in the latent space of a pretrained variational autoencoder rather than directly in pixel space (Rombach et al., 2022). Given an image x , a VAE encoder first maps it to a latent representation $z_0 = \mathcal{E}(x)$, after which the forward and reverse diffusion processes operate on z_t instead of x_t . Text-to-image diffusion models further condition

the denoising network on a prompt: a text encoder τ maps a caption c to conditioning tokens $\tau(c)$, which are supplied to the denoiser through cross-attention. GENDA keeps the VAE, text encoder, and denoiser frozen and replaces this text-only conditioning sequence with the joint visual-text sequence $[h_\psi(\tilde{v}); \tau(c)]$ described in Section 3.3.

B Implementation Details

All experiments were conducted on $2 \times$ NVIDIA A100 GPUs (40 GB each). We adopt DeepSpeed ZeRO-2, gradient accumulation and gradient checkpointing to save memory for every run. Under this configuration, training cost comparison between two methods is provided in Table 5.

Table 5: Efficiency and Cost Quantification. Added parameters are relative to LLaVA-1.5.

	LLaVA-1.5	GENDA
Stage-1 wall-clock	22 h	38 h
Stage-2 wall-clock	21 h	21 h
End-to-end	43 h	59 h
GPU memory	26 GB	37 GB
Added frozen parameters	–	1.07B
Added trainable parameters	–	0.8M

For optimization hyperparameters, unless otherwise noted, we follow the LLaVA-1.5 recipe (Liu et al., 2024a). The complete configuration we use is summarized in Table 6.

Table 6: Optimization hyperparameters used for GENDA Stage-1 and Stage-2 training.

Hyperparameter	Stage 1	Stage 2
Trainable modules	f_θ, g_ϕ, h_ψ	g_ϕ, LLM
Optimizer	AdamW	AdamW
Learning rate (LLM)	–	2×10^{-5}
Learning rate (proj.)	1×10^{-3}	2×10^{-5}
Learning rate (f_θ)	2×10^{-5}	–
Weight decay	0.0	0.0
Warmup ratio	0.03	0.03
LR schedule	cosine	cosine
Effective batch size	256	128

C Datasets

C.1 Training Data

Stage-1 alignment uses LLaVA-558K released with LLaVA-1.5 (Liu et al., 2024a), consisting of 558,128 image-caption pairs. Stage-2 visual instruction tuning uses LLaVA-665K, the multi-modal instruction mixture from the same release. For the diffusion-backbone ablation (Section 4.4),

Table 7: Comparison of Stage-1-only (S1) and Stage-1+2 (S1+S2) generative supervision.

Method	MMBench	MMStar	MMMU	MMVP	AI2D	OCRQA	GQA	TextVQA
LLaVA-1.5	64.3	33.1	35.7	24.7	54.8	58.1	62.0	58.2
+ GENDA (S1)	65.1	34.4	35.9	27.1	54.1	56.2	65.7	56.3
+ GENDA (S1+S2)	64.0	33.7	35.5	25.2	53.8	56.0	62.4	54.7

we use the COCO Karpathy training split (Karpathy and Fei-Fei, 2015) containing 113,287 images with five captions each; one caption is randomly sampled per image at each training step.

C.2 Evaluation Data

We evaluate image captioning on the following benchmarks using their standard validation/test splits, reporting CIDEr scores for all results:

- **COCO-Caps** (Lin et al., 2014): Karpathy test split (5,000 images).
- **Flickr30K** (Young et al., 2014): Karpathy test split (1,000 images).
- **nocaps** (Agrawal et al., 2019): validation split (4,500 images).
- **TextCaps** (Sidorov et al., 2020): validation split (3,166 images).
- **VizWiz-Caps** (Gurari et al., 2020): validation split (7,750 images).

Downstream zero-shot retrieval is evaluated on the COCO and Flickr30K test splits using the encoder’s image and text towers (with the corresponding text encoder originally paired with each backbone). Classification task follows the standard CLIP evaluation protocol on ImageNet-1k validation, CIFAR-10/100 test splits, and Caltech-101 with the official prompts.

For instruction-tuned VLMs, we follow each benchmark’s standard evaluation protocol:

- **MMBench** (Liu et al., 2024b): English development split, evaluated using circular-evaluation accuracy.
- **MMStar** (Chen et al., 2024): the 1,500-example public evaluation set, reporting average multiple-choice accuracy.
- **MMMU** (Yue et al., 2024): validation split, reporting overall accuracy.
- **MMVP** (Tong et al., 2024): the full 300-question benchmark comprising 150 image pairs, reporting pair-level accuracy.

- **AI2D** (Hiippala et al., 2021): test split, reporting multiple-choice accuracy.
- **OCRQA** (Mishra et al., 2019): test split, reporting exact-match answer accuracy.
- **GQA** (Hudson and Manning, 2019): test-dev split, reporting answer accuracy.
- **TextVQA** (Singh et al., 2019): validation split, reporting the standard VQA accuracy.

D Generative Instruction Tuned

We compare our Stage-1-only design with an alternative that keeps generative supervision active during Stage 2. The result is demonstrated in Table 7. Applying GENDA only shows marginal gain compared to the baseline, but still underperforms our method (Stage 1 only). Text from instruction-tuning (Stage 2) dataset are usually question-answer pairs that only focus on a very specific region of the image, and are therefore weaker conditioners compared to the image captioning used in Stage 1. We consider a proper full-Stage-2 integration a study in its own right and leave it to future work.

E Object-Density Analysis

Table 8: Image captioning performance grouped by annotated object count (# objs) on COCO and Flickr.

Object count	Images	COCO		Images	Flickr	
		LLaVA-1.5	GENDA		LLaVA-1.5	GENDA
$n \leq 4$	2,559	91.2	91.7	179	58.7	61.0
$5 \leq n \leq 8$	998	92.6	93.9	472	59.1	61.3
$n \geq 9$	1,443	90.2	97.3	349	57.3	62.6

We quantitatively test the effect of GENDA on the density of objects in images. To do this, we partition the COCO and Flickr30K test sets into three bins by annotated object count, using COCO instance annotations and Flickr30K Entities region annotations (Plummer et al., 2015). Table 8 shows that GENDA’s gain increases with scene density. This pattern shows that a short caption omits more content in dense scenes, leaving more complementary information for the generative objective to supervise.