Multimodal Understanding using Stable-Diffusion as a Task Aware Feature Extractor

Vatsal Agarwal*
University of Maryland, College Park

Gefen Kohavi
Apple

Matthew Gwilliam
University of Maryland, College Park

Eshan Verma
Apple

Daniel Ulbricht
Apple

Abhinav Shrivastava
University of Maryland, College Park

Abstract

Multimodal large language models have shown tremendous advancements in parsing and reasoning about complex scenes. However recent research has highlighted the weak vision capabilities of these models, noting that CLIP-based MLLMs fail to capture necessary vision details for the LLM to answer questions accurately. We argue that a fundamental weakness of current visual feature extraction methods is that they are unaware of the prompt and therefore cannot focus on features that are best suited for a given question. To address this, we conduct an analysis of the strength of text-to-image diffusion models and their ability to learn effective representations for multi-modal understanding. To enable task-awareness, we propose passing the prompt as input to the diffusion model. However, since these models are trained to receive captions and not questions, we design a simple instruction-tuning pipeline for efficiently finetuning diffusion models to produce question-aware image features. We highlight cases where these models excel, particularly in spatial and compositional understanding. We evaluate our approach across a variety of both general VQA and more specialized MLLM benchmarks to show the strengths and weakness of text-to-image models on visual understanding tasks, as well as provide future steps for further analysis.

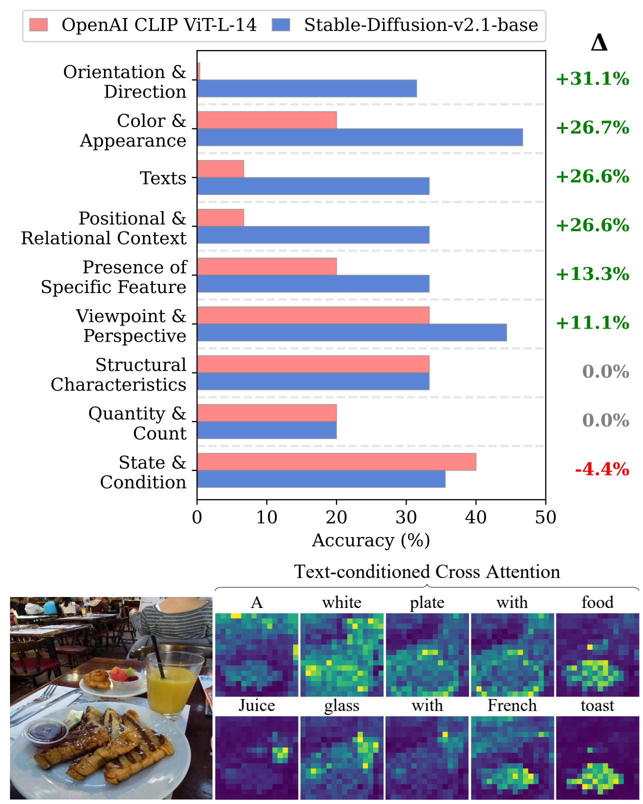

Figure 1. Text-to-image diffusion models such as Stable-Diffusion have strong capabilities for vision-language understanding. Stable-Diffusion outperforms CLIP on the challenging MMVP-VLM image-text matching benchmark [52] (shown on top), sorted by the in performance across various visual factors. Stable-Diffusion leverages its internal text-conditioned cross attention maps (shown at the bottom).

1. Introduction

Recently, there has been significant progress towards developing multi-modal large language models (MLLMs) [5, 29, 32, 33, 51]. These models rely on pre-trained vision foundation models for effective visual feature extraction and large language models (LLMs) for their advanced understanding and reasoning capabilities. To bridge the two modalities, the design of MLLMs includes a connector mechanism that projects visual information into a text space for the LLM.

*Work done during an internship at Apple.

The pairing of these two models results in enhanced multimodal understanding enabling the LLM to perform a variety of vision-language tasks such as visual question answering, image captioning, and instruction following.

Despite these advancements, these models still have many pertinent shortcomings especially related to the quality of their visual representations. Specifically, [52] found that CLIP [43], a commonly used vision encoder for MLLMs, has difficulty encoding fine-grained visual details necessary to distinguish two visually different images. This can include visual information such as orientation, structure, and viewpoint. Furthermore, [22, 30] demonstrated that these vision representations make MLLMs vulnerable to visual hallucinations such as those regarding the presence of certain objects and their quantity. Follow-up works have proposed alleviating this problem with two solutions – ensembling multiple visual encoders [23, 24, 51, 52] or incorporating extra modalities [22, 36]. However, such strategies are computationally expensive requiring more memory and increasing latency. Moreover, the choice of which vision models to use becomes a separate optimization problem with a significantly large search space [59].

We motivate that another key limitation of current multimodal models is their reliance on fixed visual features for question-answering. For instance, given a kitchen scene, in order to answer a question about what food is in a particular bowl, the visual representations must adequately localize the bowl and capture fine-grained semantic information in this region. Such dynamic information processing is also observed in how humans process visual scenes [53, 56]. As a result, these models lack flexibility in extracting relevant information for accurate question-answering. Some works have attempted to address this by developing more sophisticated modules to infuse text information into the visual features [12, 29]. However, these approaches are inefficient and only perform a form of late fusion where more granular instruction-awareness is difficult to obtain.

In contrast to both of these issues, text-to-image diffusion models have shown impressive performance in their ability to create high-quality images that capture the fine-grained semantics and compositions of the given text description [42, 44–46]. A fundamental component of these models that enables such capabilities is the cross-attention mechanism which modulates their internal activations with the input text. Examining these attention maps has shown that these generative models learn strong image-text correspondence (an example shown in Figure 1, bottom). Further works have showcased how harnessing the internal representations of these models can be used to compete with state-of-the-art models across various low and high-level vision discriminative tasks [11, 27, 41, 55].

Inspired by these works, we first explore how well off-the-shelf diffusion features perform on multi-modal understanding tasks, namely image-text matching. We perform zero-shot evaluations on various image-text benchmarks, namely Winoground [50] and MMVP-VLM [52]. We follow the protocol from He et al. [18] to extract image-text scores from the diffusion model. The results are shown in Table 1 and demonstrate that diffusion models are significantly better at capturing fine-grained details, spatial relations, and compositional information compared to CLIP and even more sophisticated CLIP variants (a comparison is also shown in Figure 1, top).

Given the promising image-text correspondence in diffusion features, we next aim to analyze how well diffusion features can capture overall image information. Specifically, we inspect model performance on the image-captioning task. Here, we follow the pre-training protocol of LLaVA [32] and train only an MLP projector layer to bridge off-the-shelf diffusion features and the pre-trained Vicuna-7B-v1.5 model [63]. We evaluate our models on the COCO-captions dataset [10, 26]. Based on this analysis, we propose leveraging text-to-image diffusion models as visual encoders for the MLLM. We investigate the performance of frozen representations first and find that while they are competitive to CLIP, they suffer because question prompts are out of distribution for the diffusion model trained with captions. Additionally, we find that the diffusion model requires more image-specific features for better grounding. As such, we propose two changes. First, during pre-training we design an implicit captioning module that leverages the CLIP image encoder to encode global image features and a trainable MLP to project them for the diffusion model. Then during the second stage of training, we propose efficiently fine-tuning the cross-attention layers of the diffusion model to better process question information when extracting image features. Together these result in a powerful image encoder capable for vision-centric multimodal understanding.

In summary, our contributions are as follows:

- We analyze the performance of off-the-shelf text-to-image diffusion features for multi-modal understanding tasks and find that they provide more granular image features than CLIP.
- We introduce a new paradigm for instruction-aware multimodal models by leveraging diffusion models as a task-aware feature extractor that can take as input the question to produce complete and relevant visual features.
- We showcase the potential of off-the-shelf text-to-image diffusion features for image-captioning and find that they aid in generating more complete and accurate captions. We identify refinement capabilities where captions can recurrently be improved over multiple passes.
- We perform comprehensive experiments on MLLM benchmarks to showcase the the benefits and drawbacks of our design compared to current state-of-the-art models, especially on vision-centric datasets.

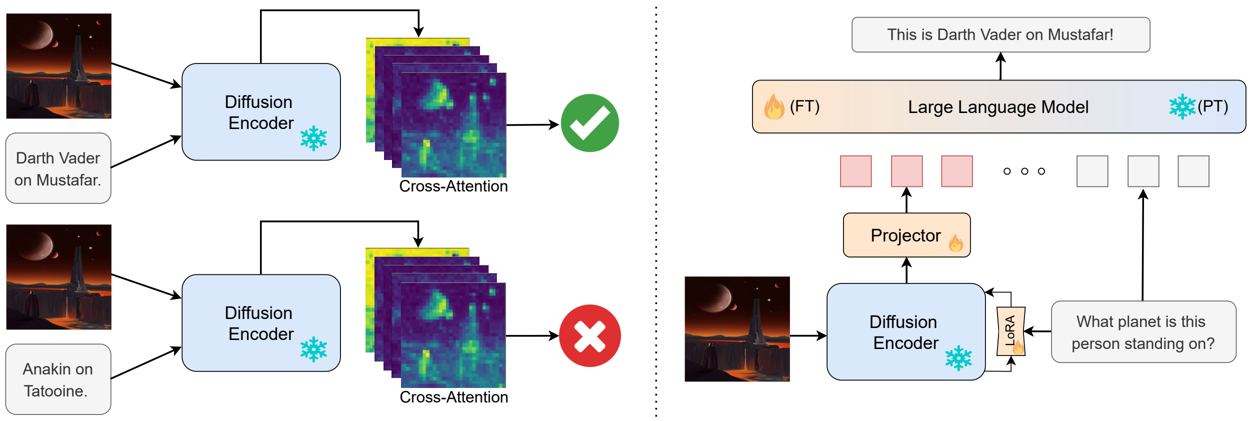

Figure 2. **Diffusion Pipelines** We have two primary setups. (**Left**) For image-text matching, we treat the text-conditioned diffusion model as a VLM, and perform the matching based on the cross attention maps. (**Right**) For captioning and question answering, given an input image and a specific question or preliminary caption, we pass both to Stable Diffusion and extract intermediate features before projecting them for the LLM. Image taken from [1].

2. Related Work

Vision Language Models Vision language modeling have been a popular topic with foundational image-text alignment papers such as [43, 60]. Multimodal LLMs go one step further and have taken the success of large scale pretrained LLMs and applied them to vision tasks [2, 12, 29, 54]. However, they typically require a large amount of pre or post training to algin the vision and language tasks. More recent methods like [9, 33, 64] show how visual instruction can be done quickly and with low data while being competitive with strong baselines [3, 29] across a wide variety of tasks.

Numerous papers and benchmarks showcase the weaknesses of these methods including their tendency for hallucination [14, 16, 21, 31, 48, 61] and general inability with spatial reasoning tasks [15, 51, 52]. Multiple improvements have been proposed such as increasing resolution [34, 37], improving data mixtures [32], as well as mixing or swapping with other encoders [20]. We show how there is still work to be done on improving vision encoders and their ability to be prompt-aware.

Combining Visual Features with Other Modalities Additional modalities have been proven useful for language tasks. [4, 39] show how a masking objective can connect more modes than RGB to language. [17, 33, 35] show how tool use alongside extra modalities can significantly expand use cases. Papers such as [7, 22, 36] show how integrating extra modalities such as depth and semantic segmentation help improve results on topics such as counting and spatial reasoning. Despite these additions, none of these models are aware of visual instruction input and therefore cannot focus features that maximize performance for a single prompt.

Diffusion Models for Discriminative Tasks There have been multiple works that look at porting diffusion models from generative tasks to discriminative tasks. The Diffusion Classifier [28] shows how to rework a standard class-conditional diffusion model into a discriminative classifier. [41] identifies where and when in a diffusion U-Net provides the strongest discriminative features.

These discriminative features have been shown to be useful for multiple tasks. For classification, [40] explores features extraction for classification and show how diffusion models are stronger than other generative models on discriminative tasks. There has also been significant explorations on using diffusion models as encoders for segmentation tasks [25, 58]. Particularly [58] shows how diffusion models have both strong open vocabulary and region-level understanding by achieving SoTA performance using a frozen diffusion backbone. Finally [18] explains that diffusion models can achieve state of the art on few-shot image-text matching. Following a similar strategy to these papers, we show how diffusion models can provide sufficiently strong discriminative features for visual instruction tuning.

3. Diffusion Preliminaries

Before delving deeper into the effectiveness of diffusion features for multi-modal understanding, we first review the underlying concepts of text-to-image diffusion models and existing feature extraction strategies.

Diffusion Models. Diffusion models are a class of generative models that aim to learn a mapping between a normal distribution and the data distribution $q(x_0)$. First, noisy samples are generated via an iterative forward noising process. At time-step $t \in [0, T]$, a noised image x_t is generated as follows:

$$x_t = \sqrt{\bar{a}_t} x_0 + (\sqrt{1 - \bar{a}_t})\epsilon \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is noise randomly sampled from a Gaus-

Table 1. Performance of various CLIP based models on different visual patterns in MMVP-VLM benchmark. Symbols are used for visual patterns due to space limit: 🧭: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, ↕: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙: Structural and Physical Characteristics, **A**: Texts, 📷: Viewpoint and Perspective. (Formatting follows [52])

Model	Image Size	Params (M)	IN-1k ZeroShot	🧭	🔍	🔄	↕	📍	🎨	⚙	**A**	📷	MMVP Average
OpenAI ViT-L-14 [43]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [43]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [62]	224^2	877.4	82.0	26.7	20.0	53.3	**40.0**	20.0	**66.7**	40.0	20.0	**53.3**	37.8
SigLIP ViT-SO-14 [62]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	**66.7**	33.3	26.7	**53.3**	37.0
DFN ViT-H-14 [13]	224^2	986.1	83.4	20.0	26.7	**73.3**	26.7	26.7	**66.7**	**46.7**	13.3	**53.3**	**39.3**
DFN ViT-H-14 [13]	378^2	986.7	**84.4**	13.3	20.0	53.3	33.3	26.7	**66.7**	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [57]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	**33.3**	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [57]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [47]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	**66.7**	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [47]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	**66.7**	26.7	20.0	33.3	33.3
SD-v2.1-base [45]	512^2	865.9	-	**31.1**	**33.3**	35.6	20.0	**33.3**	46.7	33.3	**33.3**	44.4	34.6

sian distribution. Each time-step results in an increasing amount noise such that samples from earlier time-steps are cleaner than samples from later time-steps. The amount of noise at each time-step is determined by $\{\bar{a}_t\}_{t=1}^{T}$ which is a pre-defined noise schedule.

A neural network ϵ_θ is then trained to reverse this process by learning to predict ϵ given the noisy image x_t and time-step t. For image-generation tasks, this network is most popularly uses a U-Net architecture. Thus, a trained network can take pure noise as input starting at x_T and iteratively predict ϵ to progressively generate cleaner samples $x_{T-1}, x_{T-2}, ..., x_1$ and finally x_0, representing the original data distribution. This is known as the reverse process.

For text-to-image diffusion models, ϵ_θ also takes a text condition c which is encoded via a pre-trained text-encoder $\mathcal{T}$. Thus, the noise ϵ is predicted via the updated equation:

$$f = \epsilon_\theta(x_t, t, \mathcal{T}(c)) \tag{2}$$

where $\mathcal{T}(c)$ is the encoded text description.

4. Can SD features match images and text?

In this section, we aim to understand whether features from a *frozen* Stable-Diffusion model have stronger image-understanding capabilities compared to CLIP features. To perform such an analysis in a zero-shot manner, we focus on a task that doesn't require the need for a language model, namely image-text matching. Given a set of images and text prompts, the goal of the model is to identify an image-text pair that is the most semantically aligned. We follow the protocol from [18] to generate image-text scores.

Specifically, this approach proposes using intermediate cross-attention maps between the image and text features and applying LogSumExp pooling [6] to produce a scalar value representing image-text alignment. This computation is done across all layers of the model and across multiple time-steps and the scores are then averaged to obtain the final score. For this evaluation, we compare Stable Diffusion v2.1-base [45] against CLIP-based models on two benchmarks, namely Winoground [50] and MMVP-VLM [52]. Performing well on these datasets requires the model to have strong understanding about both image semantics as well as more fine-grained details such as attributes and spatial reasoning.

The results as shown in Table 1 highlight the improved performance of the diffusion model compared to CLIP across all benchmarks. Furthermore, we observe that for the MMVP-VLM benchmark, the Stable-Diffusion model shows clear improvements in understanding orientation and direction, presence of specific features, and viewpoint and perspective patterns. Diffusion also far surpasses the other models for matched based on textual cues in the images themselves. We examine how different time-steps impact performance in Table 2 and note that features extracted from earlier time-steps result in better performance for more detailed patterns such as presence of specific features, but worse performance for color and appearance. Due to the diversity of performance across time steps, we also combine features from across a representative set of time steps – $t \in \{189, 389, 589, 789, 989\}$, and find that this strikes a good balance for decent performance across time steps. We compute our results for this ensemble as an average of 3 trials.

To complement our findings from the MMVP-VLM benchmark, we use Winoground to evaluate the diffusion models' ability to conduct compositional reasoning. This is done through a task of image-text matching with pairs of images that contain the same caption words, just in a different order. The original Winoground paper shows how

Table 2. Comparison of SD2.1 model across varying timesteps for MMVP-VLM Benchmark, using 512×512 images. For ‘Ensemble’ we use timesteps $t \in \{189, 389, 589, 789, 989\}$, and average results across 3 trials.

Model Details		MMVP-Val Benchmark									
Model	Timesteps										Avg
SD-v2.1	89	0.00	13.33	20.00	13.33	**40.00**	33.33	26.67	26.67	**46.67**	24.44
SD-v2.1	189	20.00	13.33	26.67	6.67	26.67	20.00	20.00	33.33	20.00	20.74
SD-v2.1	289	33.33	26.67	26.67	26.67	33.33	40.00	**40.00**	33.33	13.33	30.37
SD-v2.1	389	33.33	26.67	33.33	20.00	13.33	26.67	**40.00**	20.00	33.33	27.41
SD-v2.1	489	20.00	20.00	40.00	20.00	20.00	46.67	**40.00**	13.33	13.33	25.93
SD-v2.1	589	20.00	**33.33**	**53.33**	26.67	33.33	33.33	20.00	33.33	20.00	30.37
SD-v2.1	689	13.33	20.00	13.33	13.33	33.33	40.00	26.67	13.33	**46.67**	24.44
SD-v2.1	789	26.67	13.33	33.33	13.33	**40.00**	46.67	**40.00**	**40.00**	13.33	29.63
SD-v2.1	889	13.33	**33.33**	33.33	**46.67**	**40.00**	60.00	33.33	26.67	26.67	**34.81**
SD-v2.1	989	**46.67**	0.00	26.67	13.33	**40.00**	**66.67**	20.00	20.00	33.33	29.63
SD-v2.1	Ensemble	31.1 ± 7.70	**33.3** ± 6.67	35.6 ± 19.25	20.0 ± 6.67	33.3 ± 13.33	46.7 ± 11.55	33.3 ± 6.67	33.3 ± 13.33	44.4 ± 7.70	34.6 ± 2.38

Table 3. Comparison of different models on the Winoground benchmark.

Model Details			Winoground Benchmark		
Model	Image Size	Timesteps	Text	Image	Group
OpenAI ViT-L-14	224	n/a	27.75	7.75	11.75
OpenAI ViT-L-14	336	n/a	28.50	8.25	11.25
SigLIP ViT-SO-14	224	n/a	11.75	1.25	6.50
SigLIP ViT-SO-14	384	n/a	17.50	4.25	11.00
DFN ViT-H-14	224	n/a	**38.50**	11.50	14.25
DFN ViT-H-14	378	n/a	**38.50**	13.25	**15.25**
MetaCLIP ViT-L-14	224	n/a	32.50	10.75	**15.25**
MetaCLIP ViT-H-14	224	n/a	34.25	11.00	15.25
EVA01 ViT-g-14	224	n/a	27.25	9.25	11.25
EVA02 ViT-bigE-14+	224	n/a	32.00	10.50	13.50
Stable-Diffusion-v2.1-base	512	[189, 389, 589, 789, 989]	31.92 ± 2.65	**14.17** ± 1.15	10.50 ± 1.09

popular models struggle for this type of reasoning [50]. Our results, in Table 3, show slight improvements when comparing to strong CLIP baselines across text and image matching, although we do not surpass more recent models like DFN-CLIP [13] and EVA-CLIP [47]. For an ablation demonstrating the impact of time steps and noise sampling on stable diffusion performance, see the Appendix.

5. Can SD features describe an images?

Feature Extraction from Diffusion Models In this work, we use Stable Diffusion as our text-to-image diffusion model. Specifically, we use SDXL [42] as our base model given its impressive generative capabilities. There have been several explorations of how to best extract features from these diffusion models for different discriminative tasks [38, 49, 58]. These approaches generally perform a single forward pass through the U-Net to extract relevant image features. The two primary considerations for feature extraction are the choice of time-step and the choice of layers from which to extract features. We follow the latest literature [38] to choose layers for feature extraction and examine the choice of time-steps across various tasks to better determine which time-step is most optimal.

Following the LLaVA [33], we leverage these features in the pipeline shown in Figure 2 to generate captions for images. We focus our investigation on how the output captions vary depending on how we treat the text condition. That is, we experiment with various classifier-free guidance scales, as well as with different text inputs to the diffusion model, at both train and test time.

We examine model performance at both stages of LLaVA training. Specifically, the first stage trains a lightweight projection layer that is able to convert visual features into a representation that the LLM understands (PT). We additionally investigate the fully-tuned setting where the projector layer and the LLM are jointly fine-tuned on instruction-following data and use the LLaVA-Mix665k dataset [32] (FT, full-tuning). It is important to note that this fully-tuned version does not train with standard captioning setups (e.g. COCO-Captions [10] and is instead trained on instruction-

Table 4. Comparison of models on the COCO-Captions Benchmark. 512×512 images for SDXL, 336×336 images for CLIP.

Model Details			COCO-Captions Benchmark			
Model	Train Mode	Val Mode	ROUGE-L	CIDEr	B@4	SPICE
Stable-Diffusion-XL-base (PT)	No Captions	No Captions	37.11	25.80	10.90	15.65
Stable-Diffusion-XL-base (PT)	No Captions	GT Captions	37.33	25.92	11.02	15.72
Stable-Diffusion-XL-base (PT)	CFG=1.5	No Captions	31.28	21.93	8.06	11.72
Stable-Diffusion-XL-base (PT)	CFG=1.5	Pseudo-Captions	31.39	20.25	7.86	12.40
Stable-Diffusion-XL-base (PT)	CFG=1.5	GT Captions	46.97	59.16	19.63	21.66
PT-LLaVA	No Captions	No Captions	38.61	37.25	11.52	20.58
Stable-Diffusion-XL-base (PT)	CFG=1.5	PT-LLaVA-Captions	38.89	33.98	12.58	18.91
Stable-Diffusion-XL-base (PT)	CFG=1.5 w/ 30% caption dropout	PT-LLaVA-Captions w/ CFG=1.5	38.89	32.42	12.48	19.02
Stable-Diffusion-XL-base (FT)	CFG=1.5 w/ 30% caption dropout	PT-LLaVA-Captions w/ CFG=1	50.45	78.28	24.95	21.87
FT-LLaVA	No Captions	No Captions	**52.28**	**87.26**	**27.64**	**23.71**
Stable-Diffusion-XL-base (PT)	CFG=1.5	FT-LLaVA-Captions	45.62	55.18	17.86	20.50
Stable-Diffusion-XL-base (PT)	CFG=1.5 w/ 30% caption dropout	FT-LLaVA-Captions w/ CFG=1.5	44.55	49.17	16.79	20.25
Stable-Diffusion-XL-base (FT)	CFG=1.5 w/ 30% caption dropout	FT-LLaVA Captions w/ CFG=1.5	50.87	80.63	25.61	22.26

following data and then evaluated for captioning.

Table 4 shows our main results. While the best results are with the fine-tuned Llava with CLIP feature extractor (FT-LLaVA), we highlight some interesting findings for stable diffusion.

Diffusion Models as Vision Backbones First, stable diffusion features are more informative for captioning when some text is provided. In fact, even when we train the projection layer with no text inputs to stable diffusion (Train Mode "No Captions"), the captioning results are still better if we give some captions at test time (Val Mode "GT Captions"). Obviously "GT Captions" is not a fair evaluation setting, since SDXL receives the captioning targets as its input. However, is meant primarily as an oracle, to show the potential positive impact of text.

Second, stable diffusion is capable of guiding LLaVA to improve upon its initial text inputs. This can be seen when passing the PT-LLaVA captions to SDXL, as with the Val Mode "PT-LLaVA-Captions" both with and without CFG. In fact, when we finetune the SDXL, we get results that are much more competitive with the "FT-LLaVA." However, as the "FT-LLaVA-Captions" Val Mode results show, "FF-LLaVA" itself still acts as a sort of upper bound, even when we finetune the LLM for the SDXL-backbone VLLM as well.

Third, from an ablation in Table 5, we find the ideal CFG for a fair evaluation setting is on the lower end (1.5). We try to further understand the impact of CFG qualitative by computing PCA maps on SDXL features in Figure 3. The text-conditioned features are generally more semantically structured than the unconditional features. When we perform the CFG computation (weighted subtraction of undcoditional from text-conditioned features), these semantics are further emphasized. However, the ideal CFG for the oracle setting ("GT Captions") is the opposite (4.5 is ideal, but even 7.0 is better than 1.5). Since the higher CFG should further align the features to the text inputs, the inverse trend indicated that the text might be directly "leaking" from the SDXL inputs to the LLaVA outputs, particularly with higher CFG.

Table 5. Comparison of models on the COCO-Captions Benchmark. 512 × 512 images for SDXL, exploration of impact of classifier-free guidance (CFG), training only the projection module with ground truth captions and the indicated CFG.

Model Details			COCO-Captions Benchmark			
Model	CFG	Val Mode	ROUGE-L	CIDEr	B@4	SPICE
SDXL	1.5	PT-LLaVA-Captions	38.89	33.98	12.58	18.91
SDXL	1.5	GT Captions	46.97	59.16	19.63	21.66
SDXL	4.5	PT-LLaVA-Captions	36.59	28.84	9.91	19.02
SDXL	4.5	GT Captions	**52.67**	**76.05**	**24.65**	**24.44**
SDXL	7.0	PT-LLaVA-Captions	34.07	21.63	8.28	18.18
SDXL	7.0	GT Captions	49.38	62.00	20.22	23.41

Fourth, we investigate this leaking phenomenon directly to find, in Table 6, that the captions do in fact leak when using CFG. We set up an experiment where the SDXL-based MLLM is trained as normal (indicated in "Train Mode"). However, for the evaluation, the model is passed pairs of images, and captions that do not match those images. Instead of computing captioning metrics for the matching captions, we compute metrics with the mismatching captions that were used as input to the SDXL. If there is no leaking, the models should perform poorly, since we are evaluating them against text that does not match the images. On the other hand, if there is leaking they should perform well. As the Table 6 shows, while without CFG there is very limited leaking, this completely changes for higher values of CFG.

Finally, we find that by training with some caption dropout (Train Mode "30% caption dropout"), we not only get the best results shown in Table 4, we also mitigate the leaking in Table 6. With dropout, the model is as bad as no CFG for the "Mismatched" captioning (meaning it does

Table 6. Comparison of models on the COCO-Captions Benchmark. 512×512 images for SDXL, exploration of the impact of classifier-free guidance (CFG) on leaking of text from the stable diffusion inputs to the LLM outputs. To accomplish this, we evaluate in a "Mismatched" setting, where we sample a given image and an unrelated caption for input to the SDXL. We use features from SDXL to compute captions. We then compute captioning metrics relative to the unrelated input captions. If the metrics are "good," this means the SDXL leaks text to the point where the LLM hallucinates content unrelated to the image.

Model Details		COCO-Captions Benchmark			
Model	Train Mode	ROUGE-L	CIDEr	B@4	SPICE
SDXL	GT Captions, No CFG	29.25	7.32	4.15	4.64
SDXL	CFG=1.5	36.03	21.39	8.99	9.78
SDXL	CFG=1.5 w/ 30% caption dropout	30.96	10.39	5.30	6.04
SDXL	CFG=4.5	49.58	63.48	20.32	21.43

not leak severely), but on par with CFG=1.5 in the standard setting. Thus, the dropout training clearly helps the model learn a better balance between extracting image features, and simply learning to decode the text information present in SDXL features. This also somewhat aligns with how the SDXL model is trained with CFG in the first place, sometimes masking the caption for better performance.

6. Can SD features answer hard questions?

6.1. Model Design

In this section, we endeavor to understand if text-to-image diffusion models can extend beyond just describing an image, but also be leveraged as a tool for accurate instruction-following. Specifically, we start with our existing pipeline from Section 5 and make one fundamental change. Rather than feeding captions as the text-prompt for the diffusion model, we instead pass instructions as shown in Fig 2. Through this, we exploit the powerful image-text correspondence in the network's cross-attention layers to effectively focus on instruction-specific regions and features.

However, this network is not initially trained to take instructions as input and therefore alignment is necessary. Furthermore, during training, multiple instructions are often stacked together in the same conversation. This results in some instructions exceeding the token length of the diffusion text-encoder. For the purpose of our analysis, we propose two simple strategies to address each of these issues during the second-stage of LLaVA training (supervised fine-tuning, SFT).

First, we propose a random question-sampling strategy. Namely, given a conversation of questions, we randomly choose a question to feed in as a prompt to the network. We additionally truncate all questions to match the token-length constraint of the diffusion encoder. This design requires the network to have learned how to deal with imperfect prompts during pre-training and as such we experiment with three specific strategies: passing no-captions (No-Cap.), passing the ground-truth captions (GT-Cap.), and passing noisy ground-truth captions (Noisy GT-Cap.). The latter is done via random deletion of words in the caption (e.g. 30%).

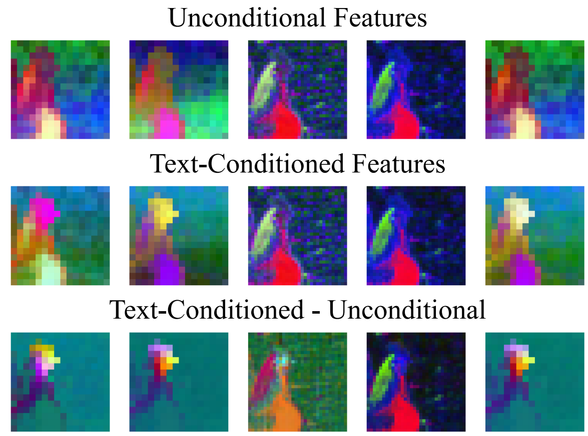

Figure 3. PCA maps of Unconditional and Text-Conditional Features. Applying CFG leads to a filtering effect on the features, where specific semantics are emphasized.

Second, to enable improved alignment of the diffusion model to instructions as text-prompts, we propose adding LoRA [19] weights to the cross-attention layers. These layers are then updated during SFT and improve the diffusion model's robustness to noisy questions. Motivated by our previous analysis, we design an architecture to extract a combination of both conditional and unconditional features. Namely, we build SDXL-LLaVA-c which **c**oncatenates both unconditional and caption/instruction-conditioned diffusion features prior to the projection layer.

6.2. Results

We evaluate our models on a diverse set of benchmarks which test the model's ability for both instruction-following as well as visual perception. We use the LLaVA-Bench-In-the-Wild dataset for instruction-understanding, which consists of 24 images and 60 questions. This benchmark contains highly out of distribution images and asks the MLLM to answer questions that require deep world knowledge. To evaluate more fundamental visual capabilities, we use two benchmarks, MMVP [52], BLINK [15]. MMVP tests the model's spatial reasoning abilities and asks questions about orientation, color, etc. BLINK builds upon this by testing the model's ability to understand visual prompts and reason about multiple images. This benchmark covers various image properties such as semantic and functional correspondence, relative depth, etc. It is important to note that while LLaVA-Bench relies on the LLM to generate text, the vision-centric benchmarks are multiple-choice.

We first discuss the overall results as shown in Table 7. We compare with two LLaVA models, namely one trained with a frozen CLIP [43] backbone and one trained with a frozen DINO [8] backbone. For LLaVA-Bench, we ob-

Table 7. Comparison for instruction-following and vision-centric benchmarks for different training and evaluation strategies. † indicates SDXL models that received both the instruction as well as a pseudo-GT caption generated from a pre-trained CLIP-LLaVA model during evaluation. We train SDXL-based models keeping the SDXL vision-encoder frozen (SDXL) as well as with LoRA-based fine-tuning (SDXL-FT). See Sec 6 for more details.

Model Details				LLaVA-Bench-In-the-Wild				Vision-Centric Benchmarks	
Model	Backbone	PT Mode	SFT Mode	Complex	Conv	Detail	All	MMVP	BLINK
CLIP-LLaVA-v1.5-7B	ViT-L14-336	N/A	N/A	75.4	58.0	60.2	66.5	24.7	36.60
DINO-LLaVA-v1.5-7B	ViT-L14-224	N/A	N/A	62.6	37.3	38.3	48.9	22.7	34.66
SDXL-LLaVA-v1.5-7B	SDXL	No-Cap.	Instr.	52.0	35.0	29.8	41.4	22.7	35.9
SDXL-LLaVA-v1.5-7B†	SDXL	No-Cap.	Instr.	54.3	33.8	34.4	43.1	22.7	-
SDXL-LLaVA-v1.5-7B	SDXL	GT-Cap.	Instr.	54.9	33.1	22.9	40.4	17.3	36.42
SDXL-LLaVA-v1.5-7B†	SDXL	GT-Cap.	Instr.	55.6	39.3	30.3	44.4	19.3	-
SDXL-LLaVA-v1.5-7B	SDXL-FT	No-Cap.	Instr.	50.4	38.6	20.7	39.3	22.0	36.63
SDXL-LLaVA-v1.5-7B†	SDXL-FT	No-Cap.	Instr.	57.2	30.9	35.5	43.9	22.0	-
SDXL-LLaVA-v1.5-7B	SDXL-FT	GT-Cap.	Instr.	49.4	37	31.4	41.2	22.0	36.12
SDXL-LLaVA-v1.5-7B†	SDXL-FT	GT-Cap.	Instr.	58.8	36.8	34.5	46.2	22.7	-
SDXL-LLaVA-v1.5-7B-c	SDXL-FT	Noisy GT-Cap.	Instr.	56.8	34.8	21.1	41.2	21.3	36.50
SDXL-LLaVA-v1.5-7B-c†	SDXL-FT	Noisy GT-Cap.	Instr.	62.9	37.6	29.5	46.9	21.3	-

Table 8. Comparison of models on the BLINK-Val Benchmark. The table shows the performance across various tasks and modalities. We perform partial finetuning (PT) of SDXL-based models either using no captions (No-Cap.), ground truth captions (GT-Cap.), or noisy ground truth captions (Noisy GT-Cap.). We also do instruction (Instr.) finetuning (SFT) for all SDXL models.

Model Details				BLINK-Val Benchmark													
Model	Backbone	PT Mode	SFT	Sim.	Count.	Depth	Jigsaw	Art	Func. Corr.	Sem. Corr.	Spat. Rel.	Obj. Loc.	Vis. Corr.	Multi-View	Refl.	Forens.	IQ
CLIP-LLaVA-v1.5-7B	ViT-L14-336	N/A	N/A	47.41	**45.00**	**52.42**	12.00	41.03	16.26	31.65	64.84	50.00	27.33	43.61	37.31	23.48	20.00
DINO-LLaVA-v1.5-	ViT-L14-224	N/A	N/A	47.41	43.33	50	2.67	32.48	22.31	20.86	**66.43**	54.92	28.49	48.87	29.1	20.45	18
SDXL-LLaVA-v1.5-7B	SDXL	No-Cap.	Instr.	47.41	40	51.61	**12.67**	35.04	20	20.86	62.94	**57.38**	29.65	43.61	35.07	21.21	25.33
SDXL-LLaVA-v1.5-7B	SDXL	GT-Cap.	Instr.	**48.15**	39.17	51.61	5.33	42.74	23.85	20.86	64.34	53.28	25	**54.89**	36.57	22.73	21.33
SDXL-LLaVA-v1.5-7B	SDXL-FT	No-Cap.	Instr.	47.41	37.5	48.39	8.67	**48.72**	**24.62**	22.3	60.84	56.56	28.49	44.36	38.06	**24.24**	22.67
SDXL-LLaVA-v1.5-7B	SDXL-FT	GT-Cap.	Instr.	47.41	40.83	50	6.67	43.59	**24.62**	25.18	65.03	53.28	29.01	41.35	34.33	19.7	**24.67**
SDXL-LLaVA-v1.5-7B-c	SDXL-FT	Noisy GT-Cap.	Instr.	46.67	38.33	**52.42**	4.67	39.32	22.31	**32.37**	62.94	56.56	**30.23**	42.11	40.3	22.73	20

serve that the CLIP-based model is the most performant and that DINO and SDXL-based models have almost a 20 point degradation in performance. We observe that adding pseudo-captions during evaluation (as indicated by †) does noticeably improve performance with a 1.7pt improvement for SDXL-LLaVA trained without captions and a 4pt improvement for SDXL-LLaVA trained with captions. Generally, we see that pre-training with ground-truth captions (GT-Cap.) prior to SFT with instructions leads to better performance on LLaVA-Bench. Applying LoRA during SFT leads to mixed results with a small drop for SDXL-LLaVA trained with no captions and a minor bump for SDXL-LLaVA trained with GT captions. Providing these models with pseudo-captions during evaluation also demonstrates improved performance. This makes sense – captions provide the diffusion model with improved grounding for the scene, which is needed for answering complex questions.

For the MMVP benchmark, we observe that while CLIP-LLaVA has a slight advantage, DINO and SDXL-based models all achieve competitive performance. Most notably, it can be seen that adding text during evaluation does not have significant performance boosts. This could be due to the simplicity of the images in this benchmark. Finally for the BLINK benchmark, we find that all models achieve roughly the same performance at around 34-36% accuracy. To better understand more granular performance benefits of our model, we examine model performance on each category of the BLINK benchmark as shown in Table 8.

Here, we observe key trends where SDXL-based models improve over CLIP and DINO-based models. Specifically, SDXL-LLaVA models consistently improve over CLIP on functional correspondence where the goal is to identify points that are functionally similar over a set of objects. Another area where we see clear improvements with SDXL over CLIP is object localization, as each model is consistently +3 points over the CLIP baseline.

7. Conclusion

In this work, we analyze the effectiveness of diffusion features for multimodal understanding. We identify that diffusion models are able to extract features that are well-aligned to text and can capture both high-level semantics and more fine-grained details. We then propose leveraging such models as task-aware feature extractors and find that they are competitive with or exceed CLIP on vision-centric benchmarks, but degrade in performance on more general-purpose question-answering. To address this, we propose several text-prompting strategies that can substantively improve model performance across various tasks. Finally, we show that minimal fine-tuning can close the gap further between CLIP and SDXL-based models and improve overall multimodal reasoning.

References

[1] Return To Mustafar by Tessas–ART on DeviantArt — deviantart.com. https://www.deviantart.com/tessas--art/art/Return-To-Mustafar-858823901. [Accessed 15-11-2024]. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 3

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[4] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *arXiv preprint arXiv:2406.09406*, 2024. 3

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[6] Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurate computation of the log-sum-exp and softmax functions, 2019. 4

[7] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models, 2024. 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 2, 5

[11] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024. 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 3

[13] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. 4, 5

[14] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 3

[15] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 3, 7

[16] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 3

[17] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 3

[18] Xuehai He, Weixi Feng, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Discffusion: Discriminative diffusion models as few-

shot vision and language learners. *arXiv preprint arXiv:2305.10722*, 2023. 2, 3, 4

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7

[20] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3

[21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 3

[22] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002, 2024. 2, 3

[23] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Xiaopeng Zhang, Jin Li, Hongkai Xiong, and Qi Tian. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 2

[24] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024. 2

[25] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 3

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[28] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, 2023. 3

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 3

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2

[31] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 2, 3, 5

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 3, 5

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[35] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023. 3

[36] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with multi-task experts. *arXiv preprint arXiv:2303.02506*, 2023. 2, 3

[37] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 3

[38] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features, 2024. 5

[39] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[40] Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification. *arXiv preprint arXiv:2307.08702*, 2023. 3

[41] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan,

Archana Swaminathan, Tianyi Zhou, Jun Ohya, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? *arXiv preprint arXiv:2311.17921*, 2023. 2, 3

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 5

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 7

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 2

[47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 4, 5

[48] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3

[49] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[50] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 2, 4, 5

[51] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 2, 3

[52] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1, 2, 3, 4, 7

[53] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2

[54] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, pages 200–212. Curran Associates, Inc., 2021. 3

[55] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 2

[56] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6): 495–501, 2004. 2

[57] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2024. 4

[58] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023. 3, 5

[59] Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. Law of vision representation in mllms. *arXiv preprint arXiv:2408.16357*, 2024. 2

[60] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022, 2022. 3

[61] Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch:

Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023. 3

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 4

[63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023. 2

[64] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3