Towards Multimodal Understanding via Stable Diffusion as a Task-Aware Feature Extractor

Vatsal Agarwal* University of Maryland
Matthew Gwilliam University of Maryland
Gefen Kohavi Apple
Eshan Verma Apple
Daniel Ulbricht Apple
Abhinav Shrivastava University of Maryland

Abstract

Recent advances in multimodal large language models (MLLMs) have enabled image-based question-answering capabilities. However, a key limitation is the use of CLIP as the visual encoder; while it can capture coarse global information, it often can miss fine-grained details that are relevant to the input query. To address these shortcomings, this work studies whether pre-trained text-to-image diffusion models can serve as instruction-aware visual encoders. Through an analysis of their internal representations, we find diffusion features are both rich in semantics and can encode strong image-text alignment. Moreover, we find that we can leverage text conditioning to focus the model on regions relevant to the input question. We then investigate how to align these features with large language models and uncover a leakage phenomenon, where the LLM can inadvertently recover information from the original diffusion prompt. We analyze the causes of this leakage and propose a mitigation strategy. Based on these insights, we explore a simple fusion strategy that utilizes both CLIP and conditional diffusion features. We evaluate our approach on both general VQA and specialized MLLM benchmarks, demonstrating the promise of diffusion models for visual understanding, particularly in vision-centric tasks that require spatial and compositional reasoning. Our project page can be found here.

1 Introduction

Effective visual feature extraction remains a key challenge in designing robust multimodal large language models (MLLMs). Due to its vision-language pre-training, CLIP [2] is the de facto visual encoder for MLLM architectures, despite its inability to encode fine-grained visual details [1, 3, 4] and capture compositional information [5, 4]. We contend that recent attempts to address these shortcomings [6, 7, 1, 8–11] are flawed in their reliance on static visual features, which may not capture task-relevant visual details. For instance, answering a question about the visibility of a butterfly's feet, as in Fig. 1 (right), requires precise localization and fine-grained semantic understanding. Although recent methods explore question-aware visual representations [12, 13], they often require substantial architectural changes to CLIP and employ fusion at a single stage thereby limiting image-query interactions [12, 14, 15].

In this work, we leverage text-to-image diffusion models as the visual encoder (pipeline shown in Fig. 1 (left)). These generative networks can create high-quality images that align well with the semantics and compositions described by a given text prompt [16–19]. Recent studies find that this is a result of their strong internal representations as well as the cross-attention mechanism, which is embedded throughout their architecture and modulates pixel features with the input text [20–23].

*Work done during an internship at Apple

Preprint. Under review.

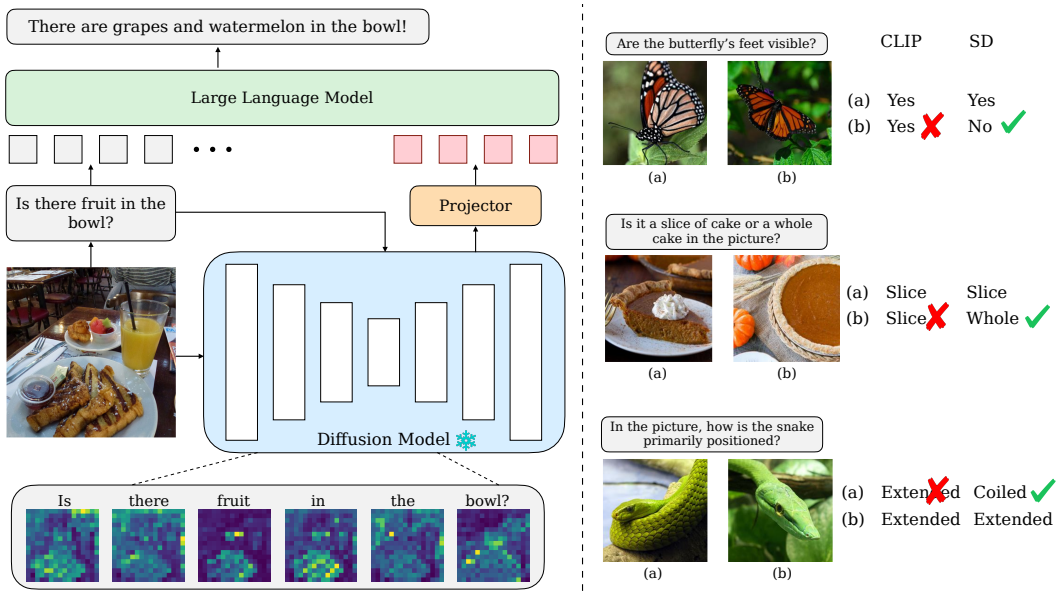

Figure 1: **Overview. (Left)** We present our full multimodal pipeline. Following LLaVA, we first extract visual features from the frozen diffusion model and pass the question as text-prompt. The LLM then uses these features to generate its answer. Cross-attention maps show that the model can use the question to focus on relevant regions **(Right)** We show examples on MMVP [1] where diffusion features outperform CLIP.

We first analyze unconditional features extracted from diffusion models across multiple blocks and timesteps. Although prior work has shown that such features can be repurposed for tasks like classification, segmentation, and depth estimation [24–27], their utility for multimodal reasoning remains underexplored. We use the LLaVA framework for our investigation and inspect model performance using features across various blocks and timesteps. Our results on general-purpose and vision-centric benchmarks show that these features encode complementary information and, in many cases, match or outperform CLIP-based LLaVA models. We show examples in Fig 1 (right).

Building on this, we inspect the text-conditional cross-attention maps. Fig. 1 (left) shows that these attention maps can focus on semantically relevant regions. We seek to quantify this alignment via image-text matching performance and find that diffusion cross-attention significantly outperforms CLIP in capturing spatial and compositional relationships (see Table 1). We also investigate how text conditioning modulates intermediate spatial features at different blocks and layers. While the impact is small for lower guidance values, amplifying this guidance can visibly change the structure of the features, with greater focus placed on image regions related to the text prompt. Furthermore, we discover a leakage effect where the language model can learn to extract the input caption fed to the diffusion model. We show how to quantify the effect of leakage and provide a mitigation strategy during training.

We then turn our focus towards using questions as the text condition for the diffusion model. First, we demonstrate that using questions as text-input guides the intermediate features to highlight relevant regions. Based on our findings, we propose leveraging the complementary information in CLIP and diffusion features to build an improved multimodal pipeline and show encouraging performance.

In summary, our contributions are as follows:

- We show that unconditional diffusion features contain fine-grained structure and semantics, leading to improved performance on vision-centric tasks with +1.23% on BLINK-val.
- We investigate what information is stored in text-conditioned cross-attention maps and find that they encode robust vision-language correspondence. We show that when used for image-text matching, these maps can match or exceed CLIP with a +5% gain on MMVP-VLM.
- We provide comprehensive analysis to show how text conditioning modulates spatial features, enabling greater focus on query-specific regions.

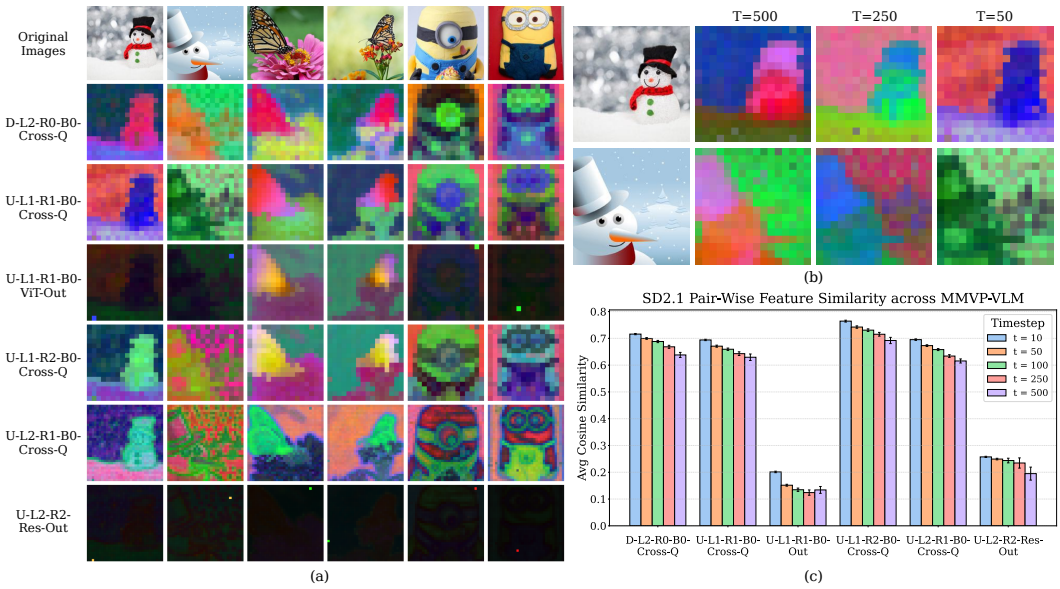

Figure 2: **Inspecting Diffusion Features**. (a, b) We visualize spatial features for three image pairs from MMVP-VLM using PCA across blocks and timesteps. For (a), we use T=50; for (b), we fix `U-L1-R1-B0-Cross-Q`. We observe: (1) different blocks capture either shared semantics or image-specific details; (2) higher timesteps encode coarse layout, while lower timesteps emphasize fine-grained structure; and (3) features like `out` and `res-out` have tokens which can act as "registers" that act as shared global descriptors across similar images. (c) We plot average cosine similarity, finding that `cross-q` representations capture greater similarity compared to `out` features.

- We explore harnessing the complementary information in CLIP and conditional diffusion features and achieve promising results with a +6% improvement on MMVP over the LLaVA-v1.5-7B baseline.

2 Related Work

Vision Language Models Vision language modeling has been a popular topic with foundational image-text alignment papers such as [28, 2]. Multimodal LLMs go one step further and have taken the success of large-scale pretrained LLMs and applied them to vision tasks [29, 30, 15, 14]. However, they typically require a large amount of pre- or post-training to align the vision and language features. More recent methods like [31–33] show how visual instruction can be done quickly with low data while being competitive with strong baselines [34, 15] across a wide variety of tasks.

Numerous works and benchmarks have demonstrated the shortcomings of these methods, including their propensity for hallucination [35–40]. Furthermore, these methods exhibit a general inability to perform spatial reasoning tasks [41, 7, 1]. Several improvements have been proposed, such as increasing the resolution [42, 43], enhancing data mixtures [44, 7], and combining or swapping with other encoders [9, 1, 7]. We demonstrate that further research is necessary to enhance vision encoders and their ability to be prompt-aware.

Combining Visual Features with Other Modalities Additional modalities have been proven useful for language tasks. [45, 46] show how a masking objective can connect more modes than RGB to language. [31, 47, 48] show how tool use alongside extra modalities can significantly expand use cases. Papers such as [11, 10, 49] show how integrating extra modalities such as depth and semantic segmentation helps 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 on 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 [50] shows

how to rework a standard class-conditional diffusion model into a discriminative classifier. [26] identifies where and when in a diffusion U-Net provides the strongest discriminative features.

Config	LLaVA-B	MMVP	BLINK-val	Natural-Bench		
	All	Acc	Acc	Q-Acc	I-Acc	G-Acc
LLaVA-v1.5-7B	66.5	24.7	36.60	37.70	43.80	14.32
Layer Configurations at T = 50						
`D-L2-R0-B0-Cross-Q`	33.4	17.3	35.81	26.71	34.87	7.11
`U-L1-R1-B0-Cross-Q`	**45.3**	**22.7**	37.28	30.23	37.05	**9.11**
`U-L1-R1-B0-Out`	41.1	22.0	**37.83**	28.82	37.05	8.68
`U-L1-R2-B0-Cross-Q`	42.2	**22.7**	34.87	**31.13**	**37.87**	8.42
`U-L2-R1-B0-Cross-Q`	39.0	22.0	37.23	28.82	35.71	8.95
`U-L2-R2-Res-Out`	34.2	19.3	35.65	27.68	35.68	6.74
Time-Steps for `U-L1-R1-B0-Cross-Q`						
T = 10	40.0	22.0	36.82	30.29	36.61	8.89
T = 50	**45.3**	**22.7**	37.28	30.23	37.05	9.11
T = 100	44.7	22.0	37.37	**30.87**	**38.45**	**9.42**
T = 250	41.3	18.0	**37.39**	26.82	34.50	7.11
T = 500	29.8	12.7	36.54	16.87	27.61	3.11

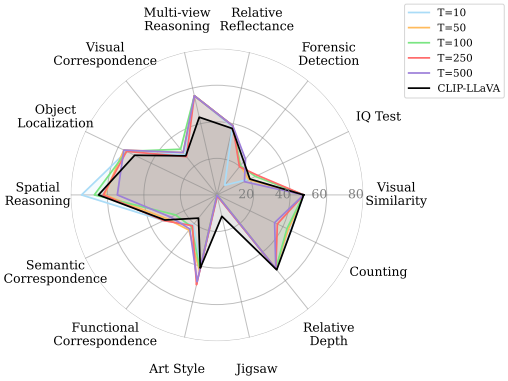

Figure 3: **General Model Performance (Left):** We evaluate multimodal reasoning using the LLaVA framework with diffusion features at different layers and timesteps. The table reports accuracy on LLaVA-Bench, MMVP, and NaturalBench under varying feature extraction points. **BLINK-val Performance (Right):** The plot shows BLINK-val benchmark performance across different timesteps. SD-based models consistently outperform CLIP (in black) across timesteps. We point out notable improvements: +10% on Spatial Reasoning, Multi-view Reasoning, and Art Style.

These discriminative features are useful for multiple tasks. For classification, [51] explores feature extraction and shows how diffusion models are stronger than other generative models on discriminative tasks. There has also been significant exploration on using diffusion models as encoders for segmentation tasks [52, 53]. Particularly [52] shows how diffusion models have both strong open vocabulary and region-level understanding by achieving SoTA performance using a frozen diffusion backbone. Finally, [5] 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 What Visual Information Do Diffusion Models Encode?

In this section, we investigate the quality of diffusion representations for visual question answering. We aim to understand how well unconditioned diffusion features align with the language space of an LLM and quantify the performance of features extracted at different blocks and timesteps. In this work, we use Stable Diffusion v2.1-base [16] as our diffusion model as it has been trained on a large and diverse set of image pairs and has shown excellent generative capabilities.

3.1 Diffusion Features Encode Semantic and Structural Information

We first analyze intermediate diffusion features via PCA building off the codebase from [54]. We aim to understand how features across layers and timesteps encode semantic and structural information. For each block and selected timestep, we extract pixel-wise features and project them onto their top three principal components, allowing us to visualize spatial patterns. These visualizations qualitatively reveal how features represent spatial layout and capture both coarse- and fine-grained structural details. We present the results in Fig. 2, and we highlight key findings below.

We primarily inspect features used in cross-attention layers, specifically the pixel-wise queries, as they most directly interact with text embeddings and are likely to be more semantically aligned. To complement this, we also examine features extracted after the cross-attention operation (`b0-out`) and after the full residual attention block (`res-out`). We use a subset of image pairs from the MMVP-VLM dataset, extracting features at timestep 50 when the noise level is relatively low.

Examining Fig. 2(a), we observe that these features encode both shared and sample-specific patterns across images. For example, in the pair of minions, PCA maps reveal consistent representations for

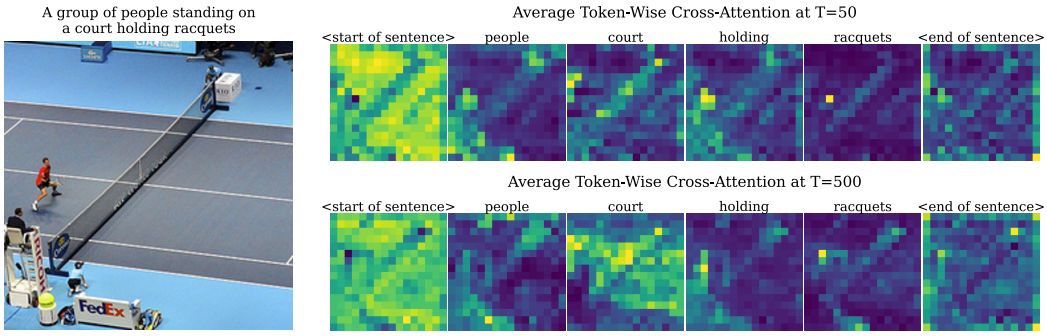

Figure 4: **Visualizing Cross-Attention Maps.** We show a sample from COCO-captions and averaged cross-attention maps at low and high timesteps for a few key words, representing the attention between pixel features and a specific word. We observe that cross-attention maps at higher timesteps show higher focus on background elements (e.g., "court"). Attention maps from lower timesteps provide improved localization of both object and action concepts (e.g., "racquets" and "holding").

object components, such as glasses and clothing. These features also capture fine-grained details, as seen in the pair of butterfly images, where the body and wing regions are clearly separated. Interestingly, features extracted after the cross-attention operation (rows 4 and 7) often encode features similar to "register" tokens: 1-2 tokens that capture the most variance in the image and are shared between samples. In Fig. 2(b), we examine how features evolve with timestep. Specifically, we extract features in a single forward pass with noise conditioned on a single timestep. We find that higher timesteps capture coarse spatial structure (e.g., the snowman's body and hat), while intermediate timesteps increasingly highlight finer details, such as the scarf (row 1, timestep 250).

Prior work by [1] identified a key limitation of CLIP: semantically similar yet visually distinct images are often embedded similarly. Motivated by this, we assess whether diffusion features better capture such visual differences. We use the MMVP-VLM benchmark, which was created by including pairs of images that had CLIP similarity over 0.95 and DINO [55] similarity below 0.6. We compute cosine similarity between feature maps of similar image pairs across blocks and timesteps as shown in Fig. 2(c). Several notable trends can be observed: (1) diffusion features capture intra-pair visual differences better compared to CLIP; (2) `cross-q` features show consistently higher pairwise similarity than `b0-out` and `res-out` features, which aligns with our earlier observations that output features encode more image-specific content; and (3) pairwise similarity decreases as the timestep increases. We hypothesize this is a result of the increased noise causing the diffusion features to encode more visually distinct features.

3.2 Diffusion Features are Effective for Vision-Centric Tasks

We investigate the effectiveness of diffusion features on multimodal understanding tasks, adopting the LLaVA architecture [31, 44] as our training framework. We begin by using off-the-shelf diffusion features and train a two-layer MLP projection head on the LLaVA-558k subset for pretraining [44]. This is followed by fine-tuning both the projection head and the language model (Vicuna-7B [56]) on the LLaVA-Mix-665k SFT dataset [44]. All training is conducted on 4 H100 GPUs, requiring approximately 4 hours for pretraining and 10 hours for fine-tuning. We defer readers to [44] for all training hyperparameters. We inspect performance across various blocks and timesteps. We evaluate using a two sets of benchmarks broadly covering instruction-following (free form generation) and visual perception (multiple choice). LLaVA-Bench-In-the-Wild (LLaVA-B) assesses instruction understanding on 24 out-of-distribution images with 60 questions requiring world knowledge. For core visual reasoning, we use MMVP [1] and NaturalBench [57] which focus on spatial reasoning (e.g.orientation, color, presence of specific features). Additionally we also use BLINK [41] which benchmarks vision-centric reasoning via visual prompting.

We show our results in Fig 3. On the left, we show model performance on all benchmarks across blocks and timesteps, and on the right, we examine more granular task-level performance on the BLINK-val benchmark. First, we observe that on LLaVA-Bench, the baseline CLIP-based LLaVA model performs the best while SD-based models have almost a 20 point degradation. For other

Table 1: **Quantifying the Quality of Diffusion Cross-Attention.** We evaluate how well diffusion cross-attention maps capture text-image alignment using the MMVP-VLM and Winoground benchmarks. We compare Stable Diffusion with various CLIP-based models across visual patterns such as viewpoint, structure, and object presence. Icons are used to denote pattern categories: 🧭: Orientation and Direction, 🔍: Specific Features, 🔄: State and Condition, 🔢: Quantity and Count, 📍: Spatial Relations, 🎨: Appearance, ⚙️: Structure, 🅰️: Text, 📷: Viewpoint. Formatting follows [1].

Model Details				**MMVP Visual Patterns**										**Winoground Benchmark**		
Model	Image Size	Params (M)	IN-1k ZeroShot	🧭	🔍	🔄	🔢	📍	🎨	⚙️	🅰️	📷	Avg.	Text	Image	Group
OpenAI ViT-L-14 [2]	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	27.75	7.75	11.75
OpenAI ViT-L-14 [2]	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	28.50	8.25	11.25
OpenCLIP ViT-H-14 [58]	224^2	986.1	78.0	20.0	13.3	60.0	33.3	13.3	53.3	40.0	6.7	26.7	29.6	30.68	11.91	8.36
SigLIP ViT-SO-14 [59]	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	11.75	1.25	6.50
SigLIP ViT-SO-14 [59]	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	17.50	4.25	11.00
DFN ViT-H-14 [60]	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**	**38.50**	11.50	14.25
DFN ViT-H-14 [60]	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	**38.50**	13.25	**15.25**
MetaCLIP ViT-L-14 [61]	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	32.50	10.75	**15.25**
MetaCLIP ViT-H-14 [61]	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	34.25	11.00	**15.25**
EVA01 ViT-g-14 [62]	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	27.25	9.25	11.25
EVA02 ViT-bigE-14+ [62]	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	32.00	10.50	13.50
SD-v2.1-base [16]	512^2	865.9	-	**31.1**	**33.3**	35.6	20.0	**33.3**	46.7	33.3	**33.3**	44.4	34.6	31.92	**14.17**	10.50

benchmarks, we observe less degradation as SD-based models approach baseline performance on MMVP with a 2 point gap. On BLINK-val, SD-based models consistently outperform CLIP models with about a 1.23% improvement. This illustrates that these features improve on pure vision-centric reasoning. On NaturalBench, we find that CLIP is able to significantly outperform SD-based models potentially due to its better vision-language alignment.

Across blocks, we observe that `cross-q` features generally perform better than `out` features such as on LLaVA-Bench and MMVP suggesting an improved alignment to text at these layers. However, we see that the feature extracted from the `down-stage` performs much worse with a 12 point drop compared to its `up-stage` equivalent on LLaVA-Bench. This aligns with findings from previous works [54], that features at the `down-stage` contain more diffusion noise. Looking at timesteps, we observe that features extracted at earlier timesteps between timestep 10-100 perform well compared to earlier or later timesteps. We hypothesize this is due to features from later timesteps losing too many fine-grained details. Inspecting BLINK-val performance (right of Fig. 3), we observe that features from different timesteps excel at different tasks. Notably, features at the earliest timestep (t=10) improve over CLIP by 10 points on spatial reasoning and match CLIP on counting. Additionally, diffusion features consistently exceed CLIP performance on multi-view reasoning regardless of timestep. Thus, while diffusion features lag behind CLIP on instruction/knowledge-heavy tasks, they offer strong advantages in pure visual reasoning, particularly when extracted from well-aligned blocks and early-to-mid timesteps.

4 How Does Text Guidance Interact with Diffusion Representations?

4.1 Cross-Attention Maps Capture Text-Aligned Visual Semantics

We first visualize internal cross-attention maps to assess how effectively diffusion models capture image-text correspondences. Using two COCO images [63] and their captions, we compute averaged cross-attention maps at a 16×16 resolution. To study the effect of noise, we extract maps at early (low) and late (high) timesteps. As shown in Fig. 4, the model accurately grounds text in visual regions—for example, "racquets" highlights the correct object. We also find that attention maps across timesteps are complementary: higher timesteps emphasize broader contexts (e.g., "court,"), while lower timesteps focus on finer details. This aligns with diffusion dynamics, where early steps model global structure and later steps refine content.

To quantify alignment, we adopt image-text matching as a proxy task, following [5]. We compute scalar alignment scores by applying LogSumExp pooling [64] over summed cross-attention maps across layers and timesteps. Based on our prior analysis, we aggregate maps from five representative timesteps ($t \in 189, 389, 589, 789, 989$) and average results across three trials. We evaluate our method on MMVP-VLM [1] and Winoground [65], two benchmarks that require fine-grained understanding of spatial relations, attributes, and compositional reasoning.

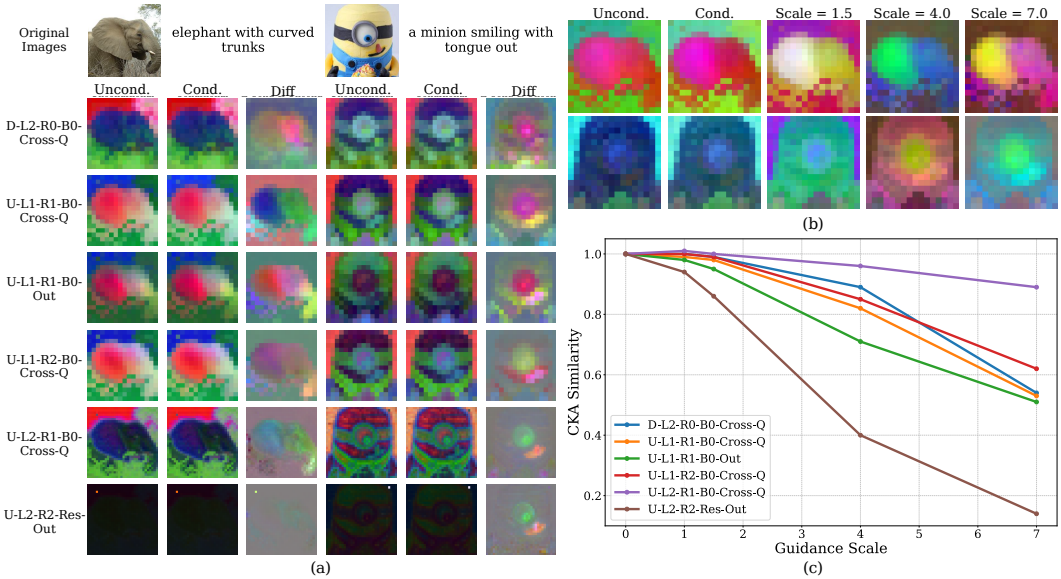

Figure 5: **Visualizing Text Conditioned Diffusion Features.** (a) A few images are sampled from the MMVP-VLM dataset and visualize PCA maps of spatial features extracted from different intermediate layers under both unconditional and text-conditioned settings (t = 50). While the overall structure of the features remains similar after adding text, we find that the difference of these features (Δ_{cond}) can highlight specific regions influenced by the text. (b) We show amplifying text guidance highlights relevant object and part regions. (c) Finally, we measure CKA similarity between unconditioned and progressively text-conditioned features across blocks and observe that spatial features extracted from `out` layers experience greater text modulation compared to `cross-q` features.

As shown in Table 1, diffusion cross-attention maps outperform CLIP-based models across all benchmarks with a 5% improvement over CLIP-ViT-H on MMVP-VLM and a 2.1% improvement on Winoground. On MMVP-VLM, diffusion excels at tasks involving orientation (+11%), identification of specific features (+20%) and viewpoint (+17%). This demonstrates that cross-attention capture robust vision-language alignments. However, diffusion does not yet surpass specialized models like DFN-CLIP [60] and EVA-CLIP [62]. Additional results, including per-timestep performance and ablations on timestep selection and noise sampling, are included in the Appendix.

4.2 Text Conditioning Modulates Spatial Feature Representations

Given the strong image-text associations observed in the cross-attention maps, we next investigate how text conditioning modulates the block-level features. We select a few images from the MMVP-VLM dataset and inspect PCA maps for both unconditional and conditional features across different blocks, with the timestep set to 50. We also visualize the PCA maps of the difference between the two features. The results are shown in Fig 5(a). We make a few key observations. First, we find that there are minimal structural changes between the unconditional and conditional features, as the PCA maps of both remain relatively similar across all blocks for all samples. However, upon inspecting the difference, we find that the change between conditional and unconditional features can highlight key attributes of the image, for instance, the elephant's ears and trunk are highlighted in the first image and the minion's tongue is localized in the second image.

To understand this effect further, we visualize how the feature changes when we amplify text-guidance. This is done via the following equation

$$X_{\text{amplified}} = X_{\text{uncond}} + s(X_{\text{cond}} - X_{\text{uncond}}) \tag{1}$$

where s is the guidance-scale and controls the amplification. We visualize the change for `U-L1-R1-B0-Cross-Q` as we increase s from 0 (purely unconditional) to 7 and show the results in Fig. 5(b). We find that while lower s settings do not modify the structure of the image, high s

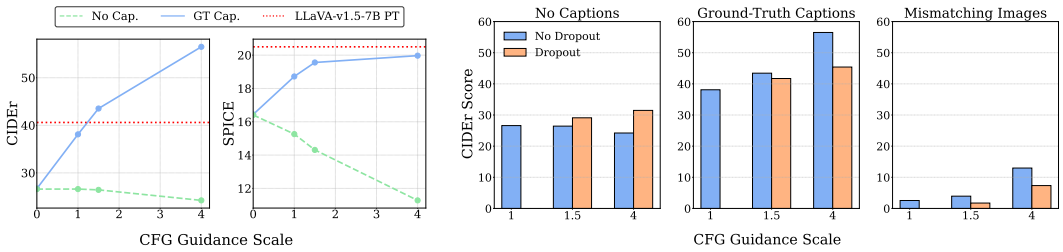

Figure 6: **COCO-Captions Performance (Left)**: We compare model performance on COCO-Captions across different pretraining text-guidance settings. Models trained with stronger ground-truth conditioning outperform the CLIP baseline when given ground-truth captions at inference but degrade significantly when no caption is provided. **Evidence of Leakage (Right):** We examine whether increasing text-guidance scale causes prompt leakage into the LLM. Training with caption dropout improves robustness when no caption is provided, with minor performance drops when ground-truth captions are used.

values increase focus on different parts o f the image such as the ears for the elephant or the mouth for the minion.

We quantify the effect of this change by performing a CKA analysis [66] between unconditional features and conditional features across various guidance scales using the COCO-captions test-set. Our intuition is that features that experience high text modulation will have lower CKA similarity with the unconditional representation as guidance-scale increases. Our results are shown in Fig. 5(c). We observe that all blocks experience substantial text modulation where `res-out` and `out` features experience the greatest change while `cross-q` features show more moderate effects.

4.3 Amplified Text Guidance Enables Leakage

Based on sensitivity to guidance-scale, we aim to empirically determine how increasing guidance changes model performance. Specifically, if increasing text-guidance can improve alignment between the diffusion features and the downstream LLM. For this analysis, we use image captioning as a proxy task and train only the projection layer on the LLaVA-558K dataset. Our intuition is that improved alignment between diffusion features and the LLM will translate to better captioning performance.

All models are trained with only the 16×16 resolution `U-L1-R1-B0-Cross-Q` feature at t = 50. We choose this layer as it exhibits good unconditional performance and also experiences reasonable text-modulation. We choose to provide the ground-truth caption as input during training as an oracle to show the potential for text-guidance. For evaluation, we use the COCO-Captions test-set and use CIDEr [67] and SPICE [68] metrics. During inference, we measure performance when passing ground-truth captions (GT Cap.) and no captions (No Cap.). The results are shown on the left of Fig 6. We observe that increasing guidance-scale during training results in improved performance when the ground-truth is provided, surpassing even CLIP. However, these models show worse performance when no caption is given. This inverse trend suggests the possibility that the LLM may actually be able to extract the input text-prompt from the diffusion features.

We investigate this phenomenon further via a mismatched setting. Namely, during inference, for an unrelated image-text pair, if the LLM is able to reconstruct the caption, then it is evidence of leakage. The results of this are shown on the right in Fig 6 (Mismatching Images) and illustrate that this is the case as the model trained with $s = 4$ is able to achieve a CIDEr score of 12.97. As a mitigation, we implement dropout during training where no caption is passed to the diffusion model randomly during pre-training. We observe that this improves robustness as captioning performance decreases on the Mismatching setting and performance on the No-Captions setting remains consistent across guidance-scales. This shows that dropout training helps the model learn a better balance between extracting image features vs. simply learning to decode the text information present in diffusion features. This also aligns with how the diffusion models are originally trained with classifier-free guidance, where the conditional prompt is occasionally masked for better generative performance.

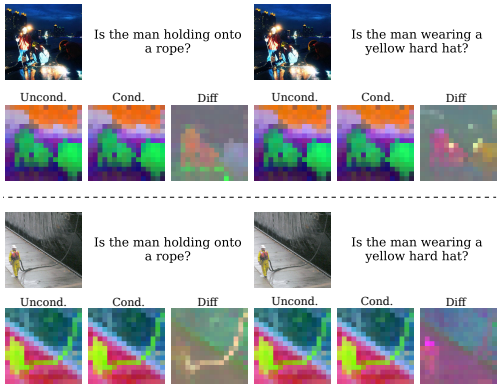

Figure 7: **Visualizing Question-Conditioned Features.** We sample an image with two questions from NaturalBench and visualize spatial features conditioned on questions via PCA. We see that the difference between conditional and unconditional features finds the relevant regions.

Config	LLaVA-B	MMVP	GQA	Natural-Bench		
	All	Acc	Acc	Q-Acc	I-Acc	G-Acc
LLaVA-v1.5-7B	66.5	24.7	62.7	37.70	43.80	14.32
SD & CLIP Concat.						
Scale = 0 (Uncond.)	64.2	26.7	63.1	39.03	44.61	13.84
Scale = 4	61	27.3	**63.2**	37.89	43.50	12.84
SD & CLIP Cross-Attn.						
Scale = 0 (Uncond.)	**66.8**	26.9	62.7	40.55	46.13	14.79
Scale = 4	63.7	**30.7**	62.5	**40.71**	**46.21**	**15.26**

Table 2: **Performance using CLIP and Conditional Diffusion Features.** We evaluate model performance when fusing CLIP and diffusion features and demonstrate improved performance when using conditional features at $s = 4$ on the MMVP and NaturalBench benchmarks.

5 Can We Extract Task-Aware Features for Question-Answering?

5.1 Conditioning with Questions Focuses on Relevant Regions

We now turn our focus to using questions as input to the diffusion model to extract question-aware features. While our previous analysis showed that amplifying text guidance with the caption increased focus on caption-related regions, we investigate whether this holds for question inputs. We sample a pair of images from the NaturalBench benchmark with corresponding questions [57]. We perform PCA analysis of the unconditional and conditional features and their difference. As shown in Fig. 7, we observe that the difference between conditional and unconditional features highlight regions that match with the question. On the top row, we see that when asking about the rope, the pipe regions are highlighted, while when asking about whether the man is wearing a yellow hat, greater focus is placed on the top of each man's head. We observe that these trends hold for the bottom image as well.

5.2 Fusing Conditional Diffusion and CLIP Improves Multimodal Understanding

Based on our analyses, we propose leveraging CLIP and conditional diffusion features for improved multimodal understanding. We experiment with two fusion strategies, namely feature concatenation and cross-attention. For cross-attention, we use the CLIP features as the queries and the diffusion features as the keys and values. To prevent the large language model from incorrectly latching onto the text, we only pass question prompts during the SFT stage of training and provide no text during the pre-training stage. We provide results with both unconditional and question-conditioned diffusion features at $s = 4$ as shown in Table 2. We find that simple concatenation with unconditional diffusion features is able to improve on MMVP and GQA with a 2 point and 0.5 point increase in performance respectively. Increasing guidance to $s = 4$, we observe further improvement. With cross-attention fusion, we observe that performance on MMVP improves by +7 points and +0.9 points on NaturalBench compared to the original LLaVA-v1.5 baseline.

6 Discussion

Broader Impacts and Limitations. Our use of Stable Diffusion v2.1 is limited to feature extraction rather than image generation, so we expect the associated risks to be substantially reduced. LLM usage also has risks with respect to hallucinating content and we advocate for careful usage of these models. While in the pipeline we study, text leaks can exacerbate the hallucination problem, we envision that an ideal ensembling strategy for leveraging each guidance scale's features could achieve better performance than CLIP and be explored as future work.

Conclusion. In this work, we investigate the potential diffusion models have as task-aware feature extractors. We show that diffusion features encode rich vision-centric information that enables improved performance on vision-centric benchmarks. We then explore how text-conditioning modulates the internal diffusion representations and showcase that its cross-attention maps capture fine-grained vision-text correspondences that propagate to block-level features. Finally, we present a simple approach to utilize the complementary information in CLIP and conditional diffusion features and showcase improved multimodal performance on general-purpose and vision-centric benchmarks.

Acknowledgements. This work was partially supported by NSF CAREER Award (#2238769). The authors would like to thank our colleagues Anubhav Gupta, Soumik Mukhopadhyay, and Pulkit Kumar for their valuable conversations and feedback. The authors acknowledge UMD's supercomputing resources made available for conducting this research. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

[1] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[3] A. K. Monsefi, K. P. Sailaja, A. Alilooee, S.-N. Lim, and R. Ramnath, "Detailclip: Detail-oriented clip for fine-grained tasks," *arXiv preprint arXiv:2409.06809*, 2024.

[4] L. Bianchi, F. Carrara, N. Messina, and F. Falchi, "Is clip the main roadblock for fine-grained open-world perception?," 2024.

[5] X. He, W. Feng, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, W. Y. Wang, and X. E. Wang, "Discffusion: Discriminative diffusion models as few-shot vision and language learners," *arXiv preprint arXiv:2305.10722*, 2023.

[6] O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, and F. Tombari, "Brave: Broadening the visual encoding of vision-language models," *arXiv preprint arXiv:2404.07204*, 2024.

[7] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," *arXiv preprint arXiv:2406.16860*, 2024.

[8] D. Jiang, Y. Liu, S. Liu, X. Zhang, J. Li, H. Xiong, and Q. Tian, "From clip to dino: Visual encoders shout in multi-modal large language models," *arXiv preprint arXiv:2310.08825*, 2023.

[9] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, "Mini-gemini: Mining the potential of multi-modality vision language models," 2024.

[10] S. Liu, L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar, "Prismer: A vision-language model with multi-task experts," *arXiv preprint arXiv:2303.02506*, 2023.

[11] J. Jain, J. Yang, and H. Shi, "Vcoder: Versatile vision encoders for multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27992–28002, 2024.

[12] R. Ganz, Y. Kittenplon, A. Aberdam, E. B. Avraham, O. Nuriel, S. Mazor, and R. Litman, "Question aware vision transformer for multimodal reasoning," 2024.

[13] R. Yu, W. Yu, and X. Wang, "Api: Attention prompting on image for large vision-language models," 2024.

[14] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[15] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, pp. 19730–19742, PMLR, 2023.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[17] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 36479–36494, Curran Associates, Inc., 2022.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022.

[19] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023.

[20] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022.

[21] R. Tang, L. Liu, A. Pandey, Z. Jiang, G. Yang, K. Kumar, P. Stenetorp, J. Lin, and F. Ture, "What the daam: Interpreting stable diffusion using cross attention," 2022.

[22] B. Liu, C. Wang, T. Cao, K. Jia, and J. Huang, "Towards understanding cross and self-attention in stable diffusion for text-guided image editing," 2024.

[23] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

[24] J. Wang, X. Li, J. Zhang, Q. Xu, Q. Zhou, Q. Yu, L. Sheng, and D. Xu, "Diffusion model is secretly a training-free open vocabulary semantic segmenter," *arXiv preprint arXiv:2309.02773*, 2023.

[25] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. 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.

[26] S. Mukhopadhyay, M. Gwilliam, Y. Yamaguchi, V. Agarwal, N. Padmanabhan, A. Swaminathan, T. Zhou, J. Ohya, and A. Shrivastava, "Do text-free diffusion models learn discriminative visual representations?," *arXiv preprint arXiv:2311.17921*, 2023.

[27] X. Chen, Z. Liu, S. Xie, and K. He, "Deconstructing denoising diffusion models for self-supervised learning," 2024.

[28] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.

[29] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 200–212, Curran Associates, Inc., 2021.

[30] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 23716–23736, Curran Associates, Inc., 2022.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[32] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[33] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[34] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[35] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023.

[36] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, "Multi-modal hallucination control by visual information grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

[37] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 18135–18143, 2024.

[38] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[39] B. Zhai, S. Yang, X. Zhao, C. Xu, S. Shen, D. Zhao, K. Keutzer, M. Li, T. Yan, and X. Fan, "Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption," *arXiv preprint arXiv:2310.01779*, 2023.

[40] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, *et al.*, "Aligning large multimodal models with factually augmented rlhf," *arXiv preprint arXiv:2309.14525*, 2023.

[41] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna, "Blink: Multimodal large language models can see but not perceive," *arXiv preprint arXiv:2404.12390*, 2024.

[42] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024.

[43] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, *et al.*, "Mm1: Methods, analysis & insights from multimodal llm pre-training," *arXiv preprint arXiv:2403.09611*, 2024.

[44] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[45] D. Mizrahi, R. Bachmann, O. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir, "4m: Massively multimodal masked modeling," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[46] R. Bachmann, O. F. Kar, D. Mizrahi, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan, and A. Zamir, "4m-21: An any-to-any vision model for tens of tasks and modalities," *arXiv preprint arXiv:2406.09406*, 2024.

[47] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "Llava-plus: Learning to use tools for creating multimodal agents," 2023.

[48] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

[49] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," 2024.

[50] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217, October 2023.

[51] S. Mukhopadhyay, M. Gwilliam, V. Agarwal, N. Padmanabhan, A. Swaminathan, S. Hegde, T. Zhou, and A. Shrivastava, "Diffusion models beat gans on image classification," *arXiv preprint arXiv:2307.08702*, 2023.

[52] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models," *arXiv preprint arXiv:2303.04803*, 2023.

[53] L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, "Diffusion models for zero-shot open-vocabulary segmentation," *arXiv preprint arXiv:2306.09316*, 2023.

[54] B. Meng, Q. Xu, Z. Wang, X. Cao, and Q. Huang, "Not all diffusion model activations have been evaluated as discriminative features," 2024.

[55] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[56] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[57] B. Li, Z. Lin, W. Peng, J. d. D. Nyandwi, D. Jiang, Z. Ma, S. Khanuja, R. Krishna, G. Neubig, and D. Ramanan, "Naturalbench: Evaluating vision-language models on natural adversarial samples," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[58] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021. If you use this software, please cite it as below.

[59] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[60] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. Toshev, and V. Shankar, "Data filtering networks," 2023.

[61] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, "Demystifying clip data," 2024.

[62] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," 2023.

[63] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.

[64] P. Blanchard, D. J. Higham, and N. J. Higham, "Accurate computation of the log-sum-exp and softmax functions," 2019.

[65] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *CVPR*, 2022.

[66] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*, pp. 3519–3529, PMLR, 2019.

[67] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[68] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398, Springer, 2016.

[69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

Towards Multimodal Understanding via Stable Diffusion as a Task-Aware Feature Extractor

Supplementary Material

A Experimental Settings

A.1 Block Configurations

We first describe our block-selection settings in more detail. We extract features at both the encoder (`down-stage`) and the decoder (`up-stage`). Additionally, we inspect mostly pixel-wise query features used in the text-conditioned cross-attention layers. We inspect these features across blocks and resolutions. Lastly, we examine a select set of output features (denoted as `out`). We adopt a structured naming convention to identify specific feature blocks within the diffusion model. Each name follows the format: `Stage-Level-Repeat-Block-FeatureType`. The stage is either D or U, followed by the resolution level (L#), the residual block index level (R#), the transformer block index (B#), and finally the type of feature extracted (Cross-Q, Out, or Res-Out). This notation allows for precise referencing of intermediate features across the model hierarchy. The full configurations are shown in Table 3.

Table 3: **Block Configurations.** We describe all block configurations used in our analysis and experiments.

Feature Config.	Res	Dim	Description
D-L2-R0-B0-Cross-Q	16×16	1280	Down-Stage, Level 2, 1st ResBlock, Cross-Attention Query
U-L1-R1-B0-Cross-Q	16×16	1280	Up-Stage, Level 1, 2nd ResBlock, Cross-Attention Query
U-L1-R1-B0-Out	16×16	1280	Up-Stage, Level 1, 2nd ResBlock, Attention Output
U-L1-R2-B0-Cross-Q	16×16	1280	Up-Stage, Level 1, 3rd ResBlock, Cross-Attention Query
U-L2-R1-B0-Cross-Q	32×32	640	Up-Stage, Level 2, 1st ResBlock, Cross-Attention Query
U-L2-R2-Res-Out	32×32	640	Up-Stage, Level 2, 3rd ResBlock, Output

Table 4: **Training Hyperparameters.** Format from [1]

.

Hyperparameter	**LLaVA-1.5**	
	Stage 1	Stage 2
batch size	256	128
lr	2e-3	2e-5
lr schedule decay	cosine	cosine
lr warmup ratio	0.03	0.03
weight decay	0	0
epoch	1	1
optimizer	AdamW [69]	
DeepSpeed stage	2	2

A.2 LLaVA Training Settings

For visual feature processing, we fix the number of tokens to 256 and resize all features to 16×16 regardless of their original resolution. Our experimental settings follow LLaVA [31, 44]. The full hyperparameters are described in Table 4. Our training protocol remains mostly unchanged, except we use DeepSpeed stage 2 for both stages of training. We train using 4 NVIDIA H100s. For our LLM, we use the Vicuna-7B [56] model.

B Full Captioning Results

We tabulate all COCO-Caption results displayed in Fig. 6 and present them in Table 5.

C Further Diffusion Feature Analysis

C.1 Visualizations

We provide more visualizations of unconditional diffusion features via the NaturalBench benchmark [57] as it consists of more complex image-pairs compared to MMVP [1]. We observe similar trends as our analysis in Sec 3.1. We observe more instances of shared objects represented similarly across each pair of images (as evidenced by the similar colors of the PCA maps). For instance, in pair (d), diffusion features across multiple blocks capture shared semantics of the motorcycle, its

Table 5: Comparison of models on the COCO-Captions benchmark. SD2.1 uses 512×512 images; CLIP uses 336×336 images.

Model Details			COCO-Captions Benchmark			
Model	Train Mode	Val Mode	ROUGE-L	CIDEr	B@4	SPICE
CLIP-ViT-L14-336	No Captions	No Captions	40.75	40.60	15.02	20.50
Stable-Diffusion-2.1-base (PT)	No Captions	No Captions	36.78	26.59	11.77	16.41
Stable-Diffusion-2.1-base (PT)	GT Captions	No Captions	34.27	26.60	11.13	15.26
Stable-Diffusion-2.1-base (PT)	GT Captions	GT Captions	39.31	38.10	14.23	18.72
Stable-Diffusion-2.1-base (PT)	GT ($s = 1.5$)	No Captions	33.43	26.43	10.53	14.31
Stable-Diffusion-2.1-base (PT)	GT ($s = 1.5$)	GT-Captions	40.70	43.53	15.46	19.56
Stable-Diffusion-2.1-base (PT)	GT ($s = 1.5$ w/ Dropout)	No Captions	36.13	29.10	11.99	16.10
Stable-Diffusion-2.1-base (PT)	GT ($s = 1.5$ w/ Dropout)	GT-Captions	40.77	41.72	15.22	19.46
Stable-Diffusion-2.1-base (PT)	GT ($s = 4$)	No Captions	31.08	24.22	9.40	11.28
Stable-Diffusion-2.1-base (PT)	GT ($s = 4$)	GT-Captions	43.74	56.50	18.10	19.97
Stable-Diffusion-2.1-base (PT)	GT ($s = 4$) w/ Dropout	No-Captions	37.52	31..48	12.49	16.10
Stable-Diffusion-2.1-base (PT)	GT ($s = 4$) w/ Dropout	GT-Captions	42.44	45.40	16.19	20.04

wheels, and the driver. Other examples include the shared encoding of the face masks in pair (e) for U-L1-R1-B0-Cross-Q and U-L1-R1-B0-Out. Interestingly, we see that this behavior is not as present in features extracted at the `down-stage`. We hypothesize that this could be a result of the diffusion noises present during the encoding stage. Lastly, we note more examples of "register" token behavior for `vit-out` and `res-out` features. This can be seen for all samples for the `res-out` features and pairs (a-d) for the `vit-out` features.

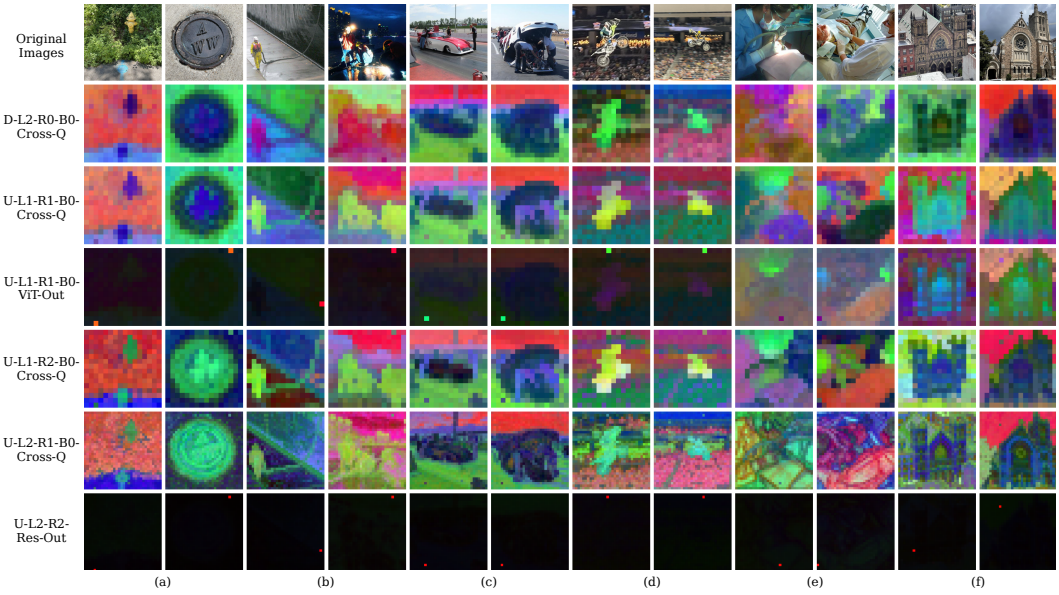

Figure 8: **More Diffusion Feature Visualizations.** We sample six pairs of images from the NaturalBench benchmark [57] and view the joint PCA maps across different blocks and layers. Please zoom in to see more details.

C.2 CKA Analysis

Here, we aim to understand the relationship between representations across different blocks. Specifically, we use CKA [66] to measure the similarity of different representations. We use the COCO-Captions test set, which consists of 5000 images, for our analysis. Furthermore, we explore how increasing text guidance may impact these block-wise relationships by computing CKA at various guidance scales, namely $s = 0, 1, 4$. As shown in Fig. 9, for $s = 0$ (purely unconditional features) `cross-q` features exhibit more similar representations compared to `out` features. When condition-

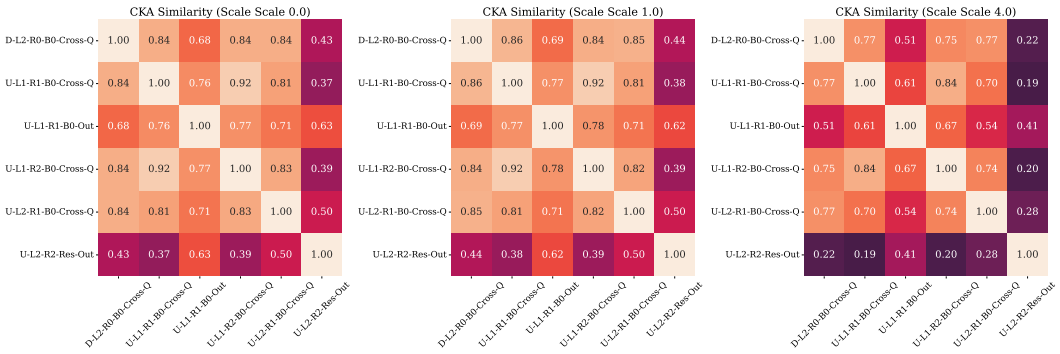

Figure 9: **CKA Block-Wise Similarity.** We compute block-wise CKA similarity using the COCO-Captions test set (5000 images) across various guidance scales ($s = 0, 1, 4$)

ing on text, we observe minimal changes in feature similarities. However, increasing text guidance ($s = 4$), we see that pair-wise CKA similarity decreases for all pairs of blocks, with a greater decrease for `out` blocks. This aligns with the trends we observed in Fig. 5 where `out` features exhibited greater changes in their representations due to higher text guidance.

D Further Cross-Attention Map Analysis

D.1 Quantifying Effect of Timesteps

Table 6: 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

We examine how different time-steps impact performance for MMVP-VLM in Table 6 and observe that features extracted from earlier timesteps yield better performance on fine-grained patterns, such as the presence of specific concepts, but worse performance for color and appearance. We do the same for the Winoground benchmark, as shown in Table 7. We find that earlier timesteps ($t \in \{89, 189, 289\}$) have a significantly lower ability to select the correct image given a caption (image-score) compared to later timesteps with a 14pt gap between features taken from t = 89 and features from t = 989. Ensembling features across multiple timesteps helps balance out performance and we observe this strategy achieves the highest text-score (ability for model to select the correct caption given an image) with relatively good performance on image- and group-score as well.

D.2 More Visualizations

Here, we provide more visualizations of cross-attention maps at both early and later timesteps. We sample images from the COCO-Captions test set to generate these maps. We observe intriguing differences in how concepts are encoded across timesteps. For example, in the skier image, later timesteps more precisely capture object-attribute bindings (e.g., the word "black" better corresponds to the coat at t = 500, compared to t = 50). We also note that early timesteps may not always be more localized than higher timesteps as the map for "coat" is slightly more concentrated at t = 500 than t = 50. For the second image of the otter, we see that the "Frisbees" map precisely localizes the frisbees

Table 7: Comparison of SD2.1 model across varying timesteps for Winoground Benchmark, using 512×512 images. For 'Ensemble' we use timesteps $t \in \{189, 389, 589, 789, 989\}$, and average results across 3 trials with 5 noise-steps.

Model Details		Winoground Benchmark		
Model	Timesteps	Text	Image	Group
Stable-Diffusion-v2.1-base (512)	89	29.25	5.75	3.00
Stable-Diffusion-v2.1-base (512)	189	23.75	8.25	4.25
Stable-Diffusion-v2.1-base (512)	289	25.50	9.25	6.50
Stable-Diffusion-v2.1-base (512)	389	26.00	12.75	10.25
Stable-Diffusion-v2.1-base (512)	489	29.50	16.00	11.25
Stable-Diffusion-v2.1-base (512)	589	29.75	15.75	10.00
Stable-Diffusion-v2.1-base (512)	689	30.25	13.75	10.50
Stable-Diffusion-v2.1-base (512)	789	27.50	17.00	12.75
Stable-Diffusion-v2.1-base (512)	889	27.50	12.50	9.00
Stable-Diffusion-v2.1-base (512)	989	28.00	**19.75**	**14.25**
Stable-Diffusion-v2.1-base (512)	[189, 389, 589, 789, 989]	**31.92** ± 2.65	14.17 ± 1.15	10.50 ± 1.09

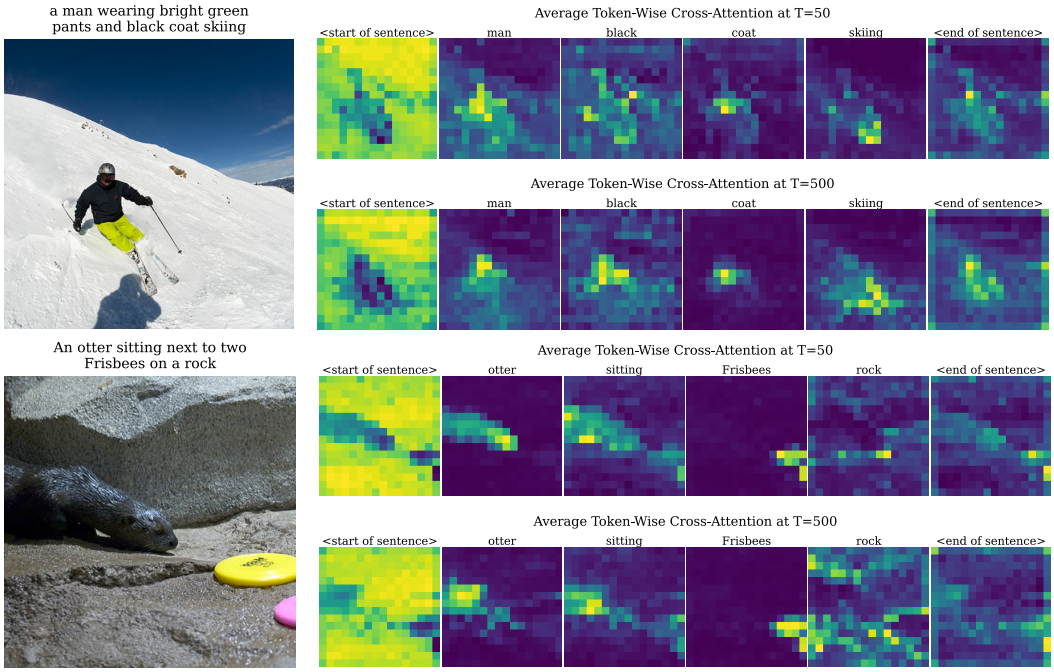

Figure 10: **More Cross-Attention Maps.** We display more examples of cross-attention maps from images in the COCO-Captions test set. For the image of the skier, we can see that object-attribute binding such as"black" and "coat" are better aligned at later timesteps compared to earlier ones.

across both timesteps. We note similar trends concerning the higher timestep maps better-capturing backgrounds, as "rock" captures the background better at t = 500, while at t = 50, the corresponding map attends more on just the edge between the rocks.

D.3 Inspecting Cross-Attention across Layers

While our earlier analysis averaged cross-attention maps across all layers at the 16×16 resolution, we now aim to understand how individual layers contribute to visual-linguistic representation. Specifically, we look at cross-attention maps at the following layers: `D-L2-R0-B0`, `D-L2-R1-B0`, `U-L1-R0-B0`, `U-L1-R1-B0`, `U-L1-R2-B0`. The maps are shown in Fig. 11. We observe that `down-stage` cross-attention maps vary considerably compared to `up-stage` maps. However,

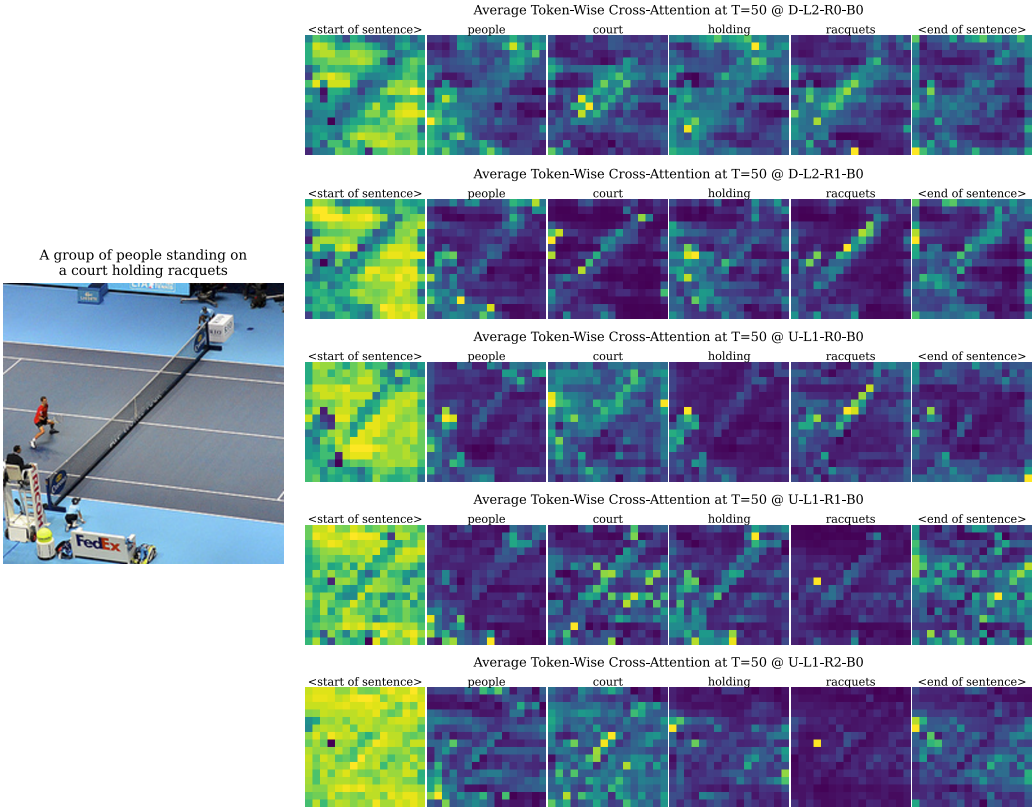

Figure 11: **Cross-Attention Maps Across Layers.** We display cross-attention maps at timestep 50 across various layers. We see that cross-attention maps are not uniform and that maps at the `up-stage` encode more robust image-text alignment.

one set of maps is not necessarily better than the others. For instance, "people" captures the person on the court in the `U-L1-R0-B0` maps, while in the `D-L2-R1-B0` maps, "people" captures the people sitting on the left side of the court. Another example is the "racquets" maps which accurately localize the object in the `U-L1-R1-B0` and `U-L1-R2-B0` maps but focus on the court net in the rest of them.

E More Examples of Question-Conditioning in Diffusion Features

In this section, we provide two more instances of how providing questions as input prompts to the diffusion model enables focus on the relevant regions. We show these in Fig. 12. To understand how question conditioning may change across resolutions, we also provide comparisons for both `U-L1-R1-B0-Cross-Q` and `U-L2-R1-B0-Cross-Q` features.

Notably, we observe that for some questions, high-resolution features can better focus on relevant regions. This is observed for the second image in the first pair of images. Specifically, for the question about the "pencil on the man's ear," we find that `U-L2-R1-B0-Cross-Q` features highlight the pencil, while `U-L1-R1-B0-Cross-Q` features instead focus on the wooden board. We also find that when the question is not aligned, the model can increase focus on the general foreground as seen in rows 3 and 4 for the `U-L1-R1-B0-Cross-Q` features. Alternatively, the model may also have a more diffuse focus or no change in focus as shown in rows 1 and 2 for the `U-L2-R1-B0-Cross-Q` features.

These examples further demonstrate the complementary information in features across different layers and resolutions. While our analysis focused on a single set of question-conditioned features, future work could explore more advanced ensembling techniques to improve model performance.

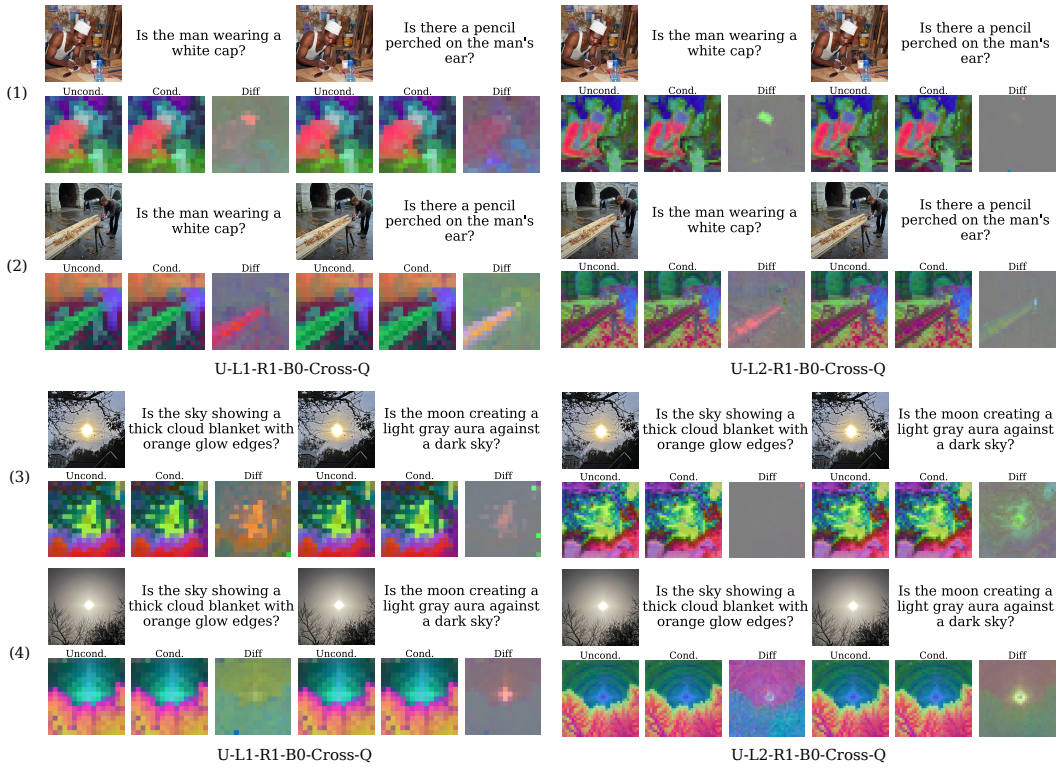

Figure 12: **More Visualizations of Question-Conditioned Features.** Please zoom in to see smaller regions of focus (e.g., pencil in second row, rightmost column of `U-L2-R1-B0-Cross-Q`)

F Investigating SDXL Architecture

Table 8: Performance Comparison of Uncond. Diffusion Features at T=50.

Model	ROUGE-L	CIDEr	BLEU @ 4	SPICE
Stable-Diffusion-2.1-base (512)	36.25	26.20	**11.54**	**16.21**
Stable-Diffusion-XL-base (512)	**37.24**	**28.30**	11.41	15.53

In this section, we investigate whether some of our key trends extend to the SDXL architecture. Namely, we focus on how well SDXL features compare to SD2.1 features and use the same protocol as Sec 4.3. We extract features from `up-level0-repeat0-vit-block7-out` from SDXL (based on analysis done in [54]). For this analysis, only the projection layer is trained, while both the vision encoder and LLM remain frozen. Our results are shown in Table 8. We observe that there are slight improvements in ROUGE and CIDEr scores when using SDXL, but SD2.1 outperforms SDXL on BLEU and SPICE.

Table 9: Measuing Leakage Effects with SDXL Features via Mismatched Setting

Model	ROUGE-L	CIDEr	BLEU @ 4	SPICE
SDXL-GT Captions ($s = 1$)	29.25	7.32	4.15	4.64
SDXL-GT Captions ($s = 1.5$)	36.03	21.39	8.99	9.78
SDXL-GT Captions ($s = 1.5$) w/ 30% caption dropout	30.96	10.39	5.30	6.04
SDXL-GT Captions ($s = 4$)	49.58	63.48	20.32	21.43

We also examine whether SDXL features exhibit similar leakage effects as SD2.1 features and perform the same comparison as in Sec 4.3 but using SDXL features. Specifically, we use the "Mismatched" setting where we pass the ground-truth caption as input to SDXL along with a randomly chosen incorrect image. As shown in Table 9, we observe that SDXL exhibits even stronger text-dependence. When trained with a guidance scale of 4, the model can more easily extract text-prompt features to reconstruct this caption even when the image is completely different. We find that adding dropout is still able to mitigate this effect.