

DiffVLM: Towards Text-Aware Multimodal Understanding via Stable Diffusion

Vatsal Agarwal*
University of Maryland
vatsalag@umd.edu

Matthew Gwilliam
University of Maryland

Eshan Verma
Apple

Daniel Ulbricht
Apple

Abhinav Shrivastava
University of Maryland

Abstract

Multimodal large language models (MLLMs) have recently achieved strong visual understanding abilities by coupling large language models with pretrained vision encoders such as CLIP. However, these models still rely on static visual representations that are computed independently of the query, limiting their ability to capture query-specific visual cues. In contrast, text-to-image diffusion models naturally produce query-aware representations through text-conditioned denoising, making them a promising complementary source of visual features. Motivated by this observation, we propose **DiffVLM**, which leverages pretrained text-to-image diffusion models to extract query-aware visual representations that complement conventional CLIP features in MLLMs. We first analyze the behavior of diffusion features, finding that they are both rich in semantics and encode strong image-text alignment. By leveraging text-conditioning, we reveal that we can steer these features to focus on regions relevant to the input question. When aligning these features with a LLM, we uncover a leakage phenomenon: amplifying text-guidance allows the LLM to reconstruct information directly from the input diffusion prompt. To address this, we introduce a mitigation strategy that preserves query-aware information while preventing leakage. We harness the complementary strengths of CLIP’s image-text alignment and diffusion features’ query-adaptive representations via a simple fusion strategy. DiffVLM improves visual question answering performance on vision-centric and general-purpose benchmarks.

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 [54] has become the de facto visual encoder in many

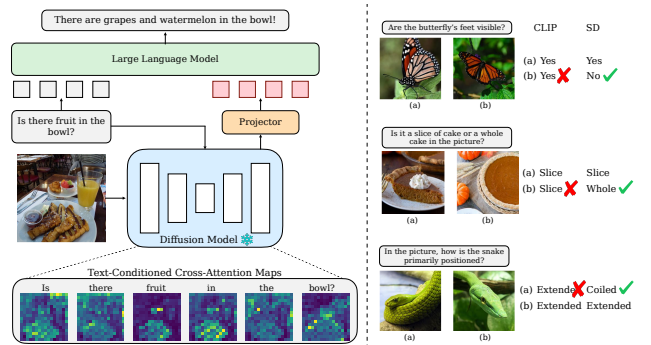


Figure 1. **Overview.** (Left) We present our text-conditioned diffusion feature-extraction pipeline. Following LLaVA, we extract visual features from the frozen diffusion model and pass the question as text-prompt. The LLM uses these features for question-answering. Cross-attention maps show that the model can use the question to focus on relevant regions (Right). We show example pairs from the MMVP benchmark [63] where diffusion features outperform CLIP.

MLLM architectures. However, its representations often fail to capture fine-grained visual details such as orientation and structure [5, 48, 63]. Moreover, they struggle to encode compositional relationships, such as the binding between objects and attributes [5, 27, 52, 74].

Recent attempts to address these shortcomings [25, 26, 28, 37, 43, 62, 63] largely rely on static visual features, which may fail to capture task-relevant visual details. For example, 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 some works explore question-aware visual representations [18, 73], they often require substantial architectural changes to CLIP and employ fusion at a single stage, thereby limiting rich interactions between the image and the query [14, 18, 35]. Finally, while [31] also proposes using diffusion features for multimodal reasoning, its focus is on analyzing how semantic information is stored in diffusion models rather than leveraging query-conditioned represen-

*Work done during an internship at Apple.

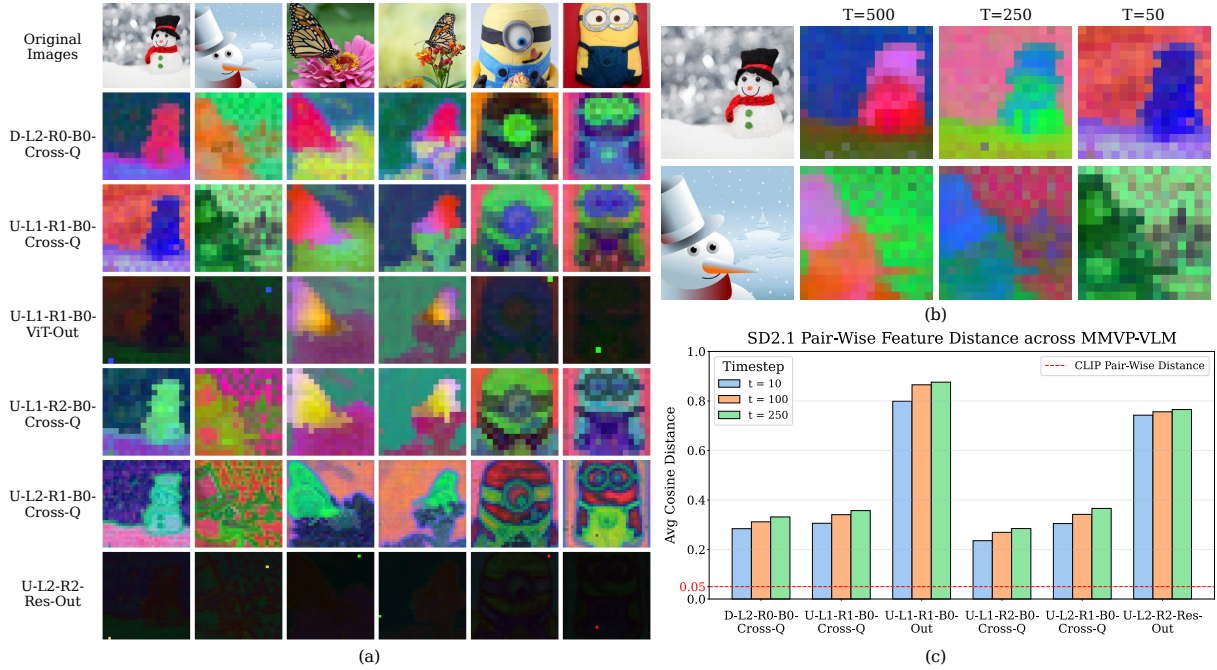


Figure 2. **Inspecting Diffusion Features.** (a, b) We visualize spatial features for three image pairs from MMVP-VLM using PCA across blocks and timesteps. For (a), we use $T=50$; for (b), we fix U-L1-R1-B0-Cross-Q. We observe: (1) different blocks capture either shared semantics or image-specific details; (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 pair-wise cosine distance of diffusion features at different blocks and timesteps over all image pairs in MMVP-VLM. We observe that diffusion features better preserve visually distinct information as all blocks exhibit pair-wise distance greater than CLIP.

tations for multimodal understanding.

In this work, we leverage text-to-image diffusion models as the visual encoder (pipeline shown in Fig. 1 (left)). These generative models can synthesize high-quality images aligned with the semantics and compositions described by a text prompt [53, 55–57]. Recent studies attribute this capability to strong internal representations and the network-wide cross-attention mechanism that modulates pixel features using text embeddings [22, 38, 60, 66]. Our key hypothesis is that these representations jointly encode general-purpose and fine-grained semantics, while textual conditioning enables query-aware modulation of visual features. Motivated by this observation, we introduce **DiffVLM**, a multimodal architecture that combines query-aware diffusion representations with the image-text aligned features of conventional vision encoders such as CLIP, leveraging their complementary strengths for improved multimodal reasoning.

We first analyze unconditional features extracted from diffusion models across multiple blocks and timesteps, using the LLaVA framework. Although prior work has shown

that such features can be repurposed for tasks such as classification, segmentation, and depth estimation [12, 30, 50, 68], their utility for multimodal reasoning remains largely unexplored. Results on both general-purpose and vision-centric benchmarks show that diffusion features encode complementary information and in many cases match or outperform CLIP-based LLaVA models (examples shown in Fig. 1 (right)).

To better understand how these features align with text, we analyze text-conditioned cross-attention maps. As illustrated in Fig. 1 (left), these maps focus on semantically relevant image regions. We quantify this alignment using image-text matching and find that diffusion cross-attention significantly outperforms CLIP in capturing spatial and compositional relationships (Table 2).

We next investigate how text conditioning modulates intermediate spatial features by manipulating its strength during feature extraction. While low guidance (lower strength) produces minimal changes, stronger guidance results in more spatially focused feature maps that highlight regions relevant to the text prompt. However, when aligning these

features to the language model, we find that at higher guidance, a leakage phenomenon occurs in which the language model can reconstruct the diffusion prompt from the features, relying on the prompt rather than the image. We quantify this effect and introduce a mitigation strategy during training.

We then leverage questions directly as the text condition, showing that this steers intermediate features toward query-relevant regions. Building on these observations, we propose a fusion strategy that combines conditional diffusion features with CLIP-style encoders to improve multimodal understanding.

In summary, our contributions are:

- **Analysis of Text-Conditioned Cross-Attention Maps.** We investigate what information is stored in text-conditioned cross-attention maps and find that they encode robust vision-language correspondence.
- **Query-Adaptive Spatial Modulation.** We show how text conditioning modulates spatial features, enabling greater focus on query-specific regions.
- **Prompt Leakage under Strong Conditioning.** We uncover a leakage phenomenon where the LLM can learn to reconstruct the diffusion prompt, quantify its extent, and propose a mitigation strategy.
- **Conditional Diffusion Feature Fusion.** We propose a simple yet effective fusion strategy combining CLIP-style vision encoders with conditional diffusion features, achieving a +6% gain over CLIP and 10% gain over SigLIP2 on MMVP.

2. Related Work

Vision Language Models Vision language modeling has been a popular topic with foundational image-text alignment papers such as [54, 72]. Multimodal LLMs go one step further and have taken the success of large-scale pre-trained LLMs and applied them to vision tasks [1, 14, 35, 65]. More recent methods like [10, 40, 77] show how visual instruction can be done quickly with low data while being competitive with strong baselines [3, 35]. Despite strong progress, numerous works and benchmarks have revealed limitations of current MLLM pipelines, including hallucination [16, 19, 23, 36, 59, 75] and weak spatial reasoning [9, 17, 62, 63, 71]. Several improvements have been proposed, such as increasing the resolution [41, 45], enhancing data mixtures [39, 62], and combining with other encoders [37, 62, 63]. These observations suggest that improving vision encoders, and enabling them to adapt to the query, remains a key challenge.

Combining Visual Features with Other Modalities Beyond RGB, prior work connects additional modalities to language via masked modeling [4, 47] or tool use [20, 42]. Works such as [8, 25, 43] show how integrating modalities such as depth and semantic segmentation improves results

on tasks such as counting and spatial reasoning. However, these models utilize static features and are unable to focus on potentially query-relevant information.

Diffusion Models for Discriminative Tasks Diffusion Classifier [33] shows how to rework a standard class-conditional diffusion model into a discriminative classifier. [50] identifies where and when in a diffusion U-Net provides the strongest discriminative features. For classification, [49] 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 [29, 70]. [21] demonstrates that diffusion models can achieve state-of-the-art on few-shot image-text matching. Building off these works, we show how diffusion models can provide sufficiently strong discriminative features for visual instruction tuning.

Revelio [31] also studies diffusion representations for multimodal understanding. Our work differs in three key aspects. First, in motivation: we propose question-conditioned diffusion features to address limitations of CLIP-style representations in multimodal reasoning, whereas Revelio focuses on analyzing how semantic information is encoded in diffusion models for general discriminative tasks. Second, our analysis provides complementary insights: we uniquely examine how diffusion features encode pairwise differences between perceptually similar images (Section 3.1), and investigate how text-conditioning impacts cross-attention mechanisms (Section 4.1) and intermediate spatial features (Section 4.2). Finally, our method leverages text-conditioning for vision-centric understanding, whereas Revelio uses unconditioned features on general-purpose benchmarks.

3. What Visual Information Do Diffusion Models Encode?

In this section, we examine what visual information is encoded in unconditional diffusion representations and their effectiveness for multimodal understanding, quantifying performance across different blocks and timesteps. We use Stable Diffusion v2.1-base [56].

3.1. Diffusion Features Encode Semantic and Structural Information

We first analyze intermediate diffusion features via PCA 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. We present the results in Fig. 2.

We are primarily interested in the query features used in the cross-attention layers, as they most directly interact with text embeddings and are likely to be more semantically

Table 1. **Unconditional Diffusion Performance.** We show performance of multimodal reasoning with diffusion features at different layers and timesteps, across LLaVA-Bench, MMVP, BLINK-val, and NaturalBench.

Vision Encoder	LLaVA-B	MMVP	BLINK-val	NaturalBench		
	All	Acc	Acc	Q-Acc	I-Acc	G-Acc
OpenAI CLIP ViT-L-14	66.5	24.7	36.6	37.7	43.8	14.3
<i>SD Layer Configurations at T = 50</i>						
D-L2-R0-B0-Cross-Q	33.4	17.3	35.8	26.7	34.9	7.1
U-L1-R1-B0-Cross-Q	45.3	22.7	37.3	30.2	37.1	9.1
U-L1-R1-B0-Out	41.1	22.0	37.8	28.8	37.1	8.7
U-L1-R2-B0-Cross-Q	42.2	22.7	34.9	31.1	37.9	8.4
U-L2-R1-B0-Cross-Q	39.0	22.0	37.2	28.8	35.7	9.0
U-L2-R2-Res-Out	34.2	19.3	35.7	27.7	35.7	6.7
<i>SD Time-Steps for U-L1-R1-B0-Cross-Q</i>						
T = 10	40.0	22.0	36.8	30.3	36.6	8.9
T = 50	45.3	22.7	37.3	30.2	37.1	9.1
T = 100	44.7	22.0	37.4	30.9	38.5	9.4
T = 250	41.3	18.0	37.4	26.8	34.5	7.1
T = 500	29.8	12.7	36.5	16.9	27.6	3.1

aligned. We also examine features extracted after the cross-attention operation (`b0-out`) and after the full residual attention block (`res-out`). We also inspect features across multiple timesteps: higher timesteps correspond to noisier, earlier stages where coarse structure dominates, while lower timesteps capture finer details as noise decreases.

Fig. 2(a) shows PCA maps of diffusion features extracted at timestep 50 (low noise) for a subset of image pairs from the MMVP-VLM dataset. These features encode both shared and sample-specific patterns across images. In the pair of minions, PCA maps reveal consistent representations for 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. Notably, features extracted after the cross-attention operation (rows 4 and 7) often contain “register” tokens: 1-2 tokens that capture the most variance in the image and are shared across pairs of images. Next, we examine how features evolve with timestep as shown in Fig. 2(b). We find that higher timesteps capture coarse spatial structure (e.g., the snowman’s body and hat), while middle timesteps increasingly highlight finer details, such as the scarf (row 1, timestep 250).

Prior work [63] 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 distance below 0.05 and DINO [51] distance above 0.4. We compute cosine distance between feature maps of similar image pairs across blocks and timesteps as shown in Fig. 2(c).

We observe several notable trends: (1) diffusion fea-

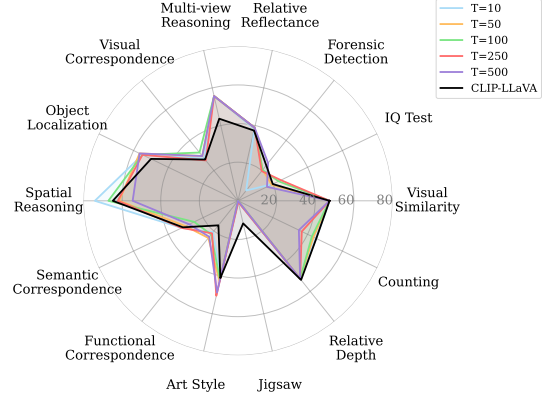


Figure 3. **BLINK-val Performance.** Benchmark performance across different timesteps. SD-based models consistently outperform CLIP (in black), with notable improvements of +10% on spatial reasoning, multi-view reasoning, and art style.

tures capture intra-pair visual differences substantially better than CLIP; (2) `b0-out` and `res-out` consistently show greater pair-wise distance than `cross-q` features, which aligns with our earlier observations that output features encode more image-specific content; and (3) pairwise distance increases as the timestep increases.

3.2. Diffusion Features are Effective for Vision-Centric Tasks

Motivated by these findings, we next investigate whether the semantics and structural information in diffusion features enable them to be effective for multimodal understanding. For this analysis, we examine the performance of unconditional diffusion features across various blocks and timesteps, using LLaVA [39, 40] as our training framework.

We use LLaVA-Bench-In-the-Wild (LLaVA-B) to assess instruction understanding on 24 out-of-distribution images with 60 questions requiring world knowledge. For core visual reasoning, we use MMVP [63] and NaturalBench [34], which test orientation, color, and other fine-grained visual attributes. We also evaluate on BLINK [17], which tests vision-centric reasoning via visual prompting.

In Table 1, we show model performance on all benchmarks across blocks and timesteps, and in Fig 3, 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. This gap narrows on MMVP, where SD-based models come within 2 points of the baseline. On NaturalBench, CLIP significantly outperforms SD-based models, again suggesting better vision-language alignment on tasks requiring broader knowledge. On BLINK-val, however, SD-based models consistently outperform CLIP by 1.2%.

Across blocks, we observe that `cross-q` features gen-

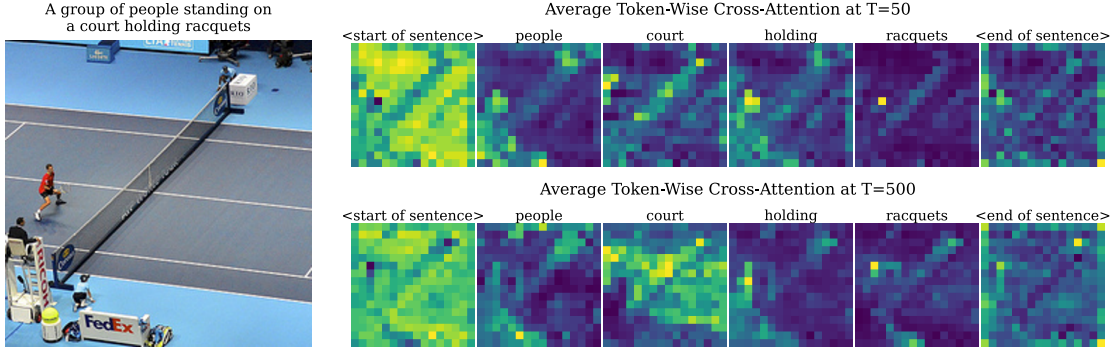


Figure 4. **Visualizing Cross-Attention Maps.** We show a sample from COCO-captions and averaged cross-attention maps at low and high timesteps for a few key words, representing the attention between pixel features and a specific word. We observe that cross-attention maps at higher timesteps show higher 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”).

erally 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 [46], where features at the `down-stage` contain more diffusion noise. Features extracted at earlier timesteps (10-100) perform better than later timesteps, likely because features from later timesteps lose too many fine-grained details. Inspecting BLINK-val performance in 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. While diffusion features lag behind CLIP on instruction-following tasks, they offer strong complementary information for pure visual reasoning.

4. How Does Text Guidance Interact with Diffusion Representations?

A key advantage of diffusion models over standard vision encoders is their ability to condition visual generation on text. In this section, we investigate how this text guidance modulates diffusion representations and whether it can be leveraged for query-adaptive visual understanding.

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 [11] and their captions, we compute averaged cross-attention maps at a 16×16 resolution. As shown in Fig. 4, the model accu-

rately grounds text in visual regions — for example, “racquets” highlights the correct object. Attention maps across timesteps are complementary: higher timesteps emphasize broader contexts (e.g., “court”), while lower timesteps focus on finer details.

To quantify alignment, we adopt image-text matching as a proxy task, following [21]. We compute scalar alignment scores by applying LogSumExp pooling [6] over summed cross-attention maps across layers and timesteps. 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 [63] and Winoground [61], which require fine-grained understanding of spatial relations, attributes, and compositional reasoning.

As shown in Table 2, 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, SD2.1 excels at tasks involving orientation (+11%), identification of specific features (+20%) and viewpoint (+17%). Notably, SD2.1 achieves this without contrastive training and even outperforms several larger CLIP variants.

4.2. Text Conditioning Modulates Spatial Feature Representations

Given the strong image-text associations observed in the cross-attention maps, we investigate how text conditioning modulates the block-level features. Fig 5(a) shows PCA maps of unconditional and conditional features and their difference across blocks. Both remain structurally similar, but the change between conditional and unconditional features highlights key attributes of the image; for instance, the elephant’s ears and trunk are accentuated 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

Table 2. **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, **A**: Text, : Viewpoint. Formatting follows [63].

Model Details				MMVP Visual Patterns										Winoground Benchmark		
Model	Image Size	Params (M)	IN-1k ZeroShot								A		Avg.	Text	Image	Group
OpenAI ViT-L-14 [54]	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 [54]	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 [24]	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 [76]	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 [76]	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 [15]	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 [15]	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 [69]	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 [69]	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 [58]	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+ [58]	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 [56]	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

done via the following equation

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

where s is the guidance-scale and controls the amplification. We visualize the change in Fig. 5(b) for U-L1-R1-B0-Cross-Q as we increase s from 0 (purely unconditional) to 7. While lower s settings do not modify the structure of the image, high s values increase focus on different parts of the image, such as the ears for the elephant or the mouth for the minion.

We quantify this change via CKA [32] between unconditional and conditional features across guidance scales on the COCO-captions test-set. Features experiencing high text modulation should show lower CKA similarity with the unconditional representation as guidance scale increases. Fig. 5(c) shows that all blocks experience substantial text modulation, with `res-out` and `out` features experiencing the greatest change and `cross-q` features showing more moderate effects.

4.3. Amplified Text Guidance Induces Leakage

Here, we examine how increasing text-guidance impacts alignment between diffusion features and the downstream LLM, using image captioning as a proxy task. Our intuition is that stronger alignment will translate to better captioning performance. For this analysis, we train only the projection layer on the LLaVA-558K dataset.

All models use U-L1-R1-B0-Cross-Q at $t = 50$ and 16×16 resolution. We provide the ground-truth caption as input during training as an oracle. For evaluation, we use the COCO-Captions test-set and use CIDEr [67] and

SPICE [2] 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. Increasing guidance-scale during training improves performance when the ground-truth is provided, surpassing even CLIP, but worsens performance when no caption is given. This inverse trend suggests that the LLM can extract the input text-prompt from the diffusion features.

We investigate this phenomenon further via a mismatched setting. Specifically, there is evidence of leakage if the LLM can reconstruct the caption for an unrelated image-text pair during inference. The results of this are shown on the right in Fig. 6 (Mismatching Images): 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, with some probability, no caption is passed to the diffusion model 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. Dropout thus helps the model balance extracting image features against decoding text information present in diffusion features, mirroring classifier-free guidance training.

5. Can We Extract Text-Aware Features for Question-Answering?

In this section, we investigate whether text-conditioning can be leveraged to extract question-aware features. While standard vision encoders produce static representations regardless of the query, diffusion models offer the opportunity to steer internal representations via conditioning on the ques-

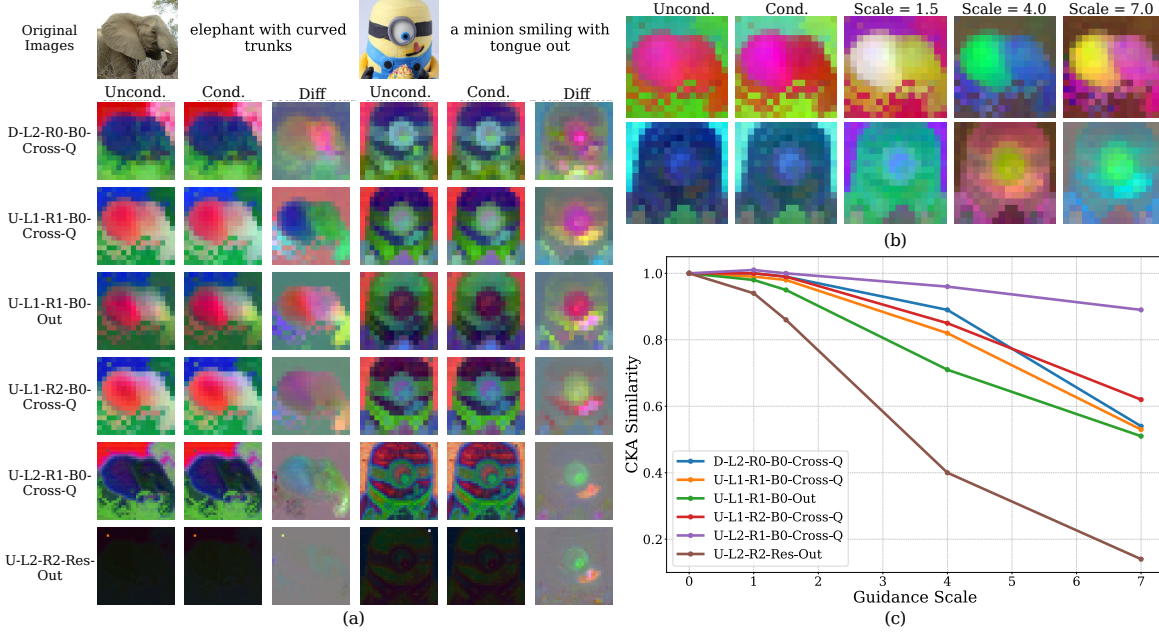


Figure 5. **Visualizing Text Conditioned Diffusion Features.** (a) A few images are sampled from the MMVP-VLM dataset and visualize PCA maps of spatial features extracted from different intermediate layers under both unconditional and text-conditioned settings ($t = 50$). While the overall structure of the features remains similar after adding text, 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 unconditional and progressively text-conditioned features across blocks and observe that spatial features extracted from `out` layers experience greater modulation compared to `cross-q` features.

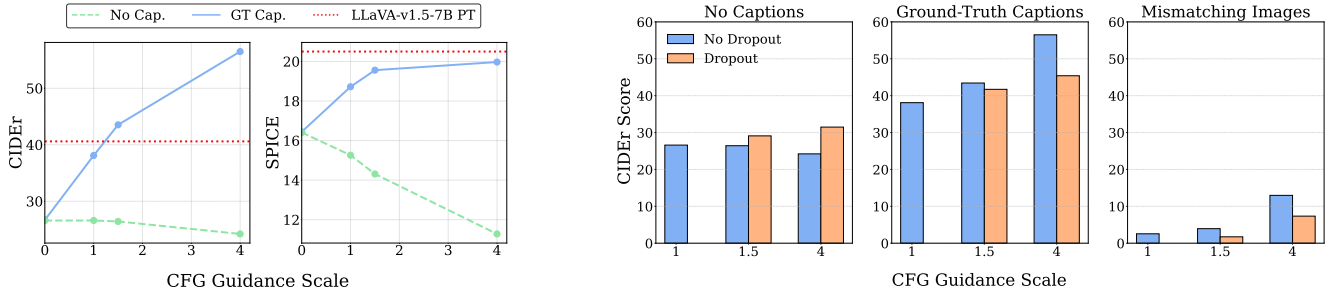


Figure 6. **COCO-Captions Performance (Left):** We compare model performance on COCO-Captions across different pretraining text-guidance settings. Models trained with stronger ground-truth conditioning outperform the CLIP baseline when given ground-truth captions at inference, but degrade significantly 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.

tion to focus on question-relevant regions. Building on these insights, we introduce DiffVLM, which harnesses the complementary strengths of CLIP and diffusion features for improved multimodal understanding.

5.1. Conditioning with Questions Focuses on Relevant Regions

We next investigate whether text-guidance improves focus on related spatial regions for question inputs, as text-to-image diffusion models are trained with captions, not ques-

tions. This is non-trivial: unlike captions, questions contain non-visual words as well as words describing visual concepts that may not actually appear in the image, making it unclear whether the model can isolate and attend to the visually relevant content.

We sample two image-question pairs from the Natural-Bench benchmark [34] and perform PCA on the unconditional and conditional features and their difference. As shown in Fig. 7, the difference highlights question-relevant regions: when asked about a rope (which is absent), the

Table 3. **Performance Fusing Conditional Diffusion Features.** We evaluate model performance when combining conditional diffusion features with conventional CLIP encoders, demonstrating improved performance on the MMVP and NaturalBench benchmarks.

Vision Encoder	Res.	LLaVA-B MMVP GQA			Natural-B		
		All	Acc	Acc	Q-Acc	I-Acc	G-Acc
CLIP ViT-L-14	336	66.5	24.7	62.7	37.7	43.8	14.3
+ SD ($s = 4$)	-	63.7	30.7	62.5	41.6	47.3	16.2
PE-Core ViT-L-14	336	62.3	33.3	63.5	42.4	47.6	16.6
+ SD ($s = 4$)	-	63.8	30.7	64.1	43.5	48.9	16.8
SigLIP2 ViT-SO-16	384	66.5	32.7	63.5	40.5	46.4	16.1
+ SD ($s = 4$)	-	67.1	42.7	64.2	44.7	50.2	18.8

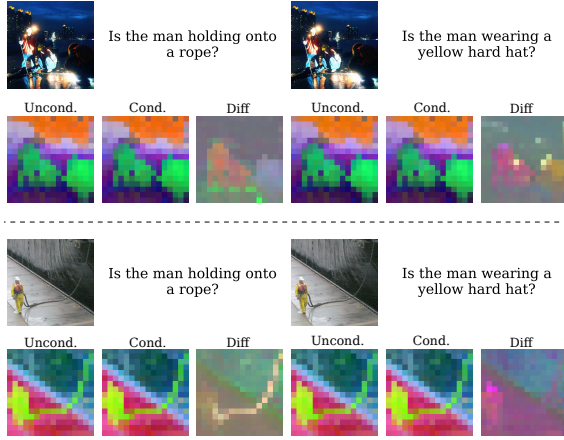


Figure 7. **Visualizing Question-Conditioned Features.** We sample an image with two questions from NaturalBench and visualize spatial features conditioned on questions via PCA. We see that the difference between conditional and unconditional features finds the relevant regions.

model attends to the pipe held by the men; when asked about a yellow hat, focus shifts to the top of each man’s head. In the bottom row, the model highlights the rope (bottom-left), and the man’s head for the ”hat” question.

5.2. Fusing Conditional Diffusion and CLIP Improves Multimodal Understanding

Motivated by our findings: (1) CLIP features provide strong image-text alignment, (2) diffusion features encode complementary vision-centric information, and (3) text-guidance enables query-adaptive feature extraction, we propose DiffVLM to integrate conditional diffusion features with CLIP encoders for enhanced multimodal understanding.

We leverage a cross-attention mechanism to fuse the fea-

Table 4. **Fusion Ablations.** We evaluate model performance when combining CLIP and diffusion features with different guidance scales and fusion strategies.

Vision Encoder	MMVP	GQA	Natural-B
	Acc	Acc	G-Acc
CLIP ViT-L-14	24.7	62.7	14.3
<i>SD & CLIP Concat.</i>			
Scale = 0 (Uncond.)	26.7	63.1	13.8
Scale = 4	27.3	63.2	12.8
<i>SD & CLIP Cross-Attn.</i>			
Scale = 0 (Uncond.)	26.9	62.7	14.8
Scale = 4	30.7	62.5	16.2

tures, where the CLIP encoder features are the queries, and diffusion features are passed as the keys and values. To prevent the large language model from incorrectly latching onto the diffusion prompt, we only pass question prompts during the SFT stage of training and provide no text during the pre-training stage. We set the guidance scale $s = 4$ for all experiments.

We evaluate our fusion strategy with CLIP, along with stronger CLIP-style encoders such as SigLIP2 [64], and Perception Encoder (PE-Core) [7]. As shown in Table 3, conditional diffusion features improve NaturalBench performance consistently across all encoders. On MMVP, we observe strong gains for CLIP (+6%) and SigLIP2 (+10%). PE-Core, which already has a stronger MMVP baseline, shows marginal improvement. Adding conditional diffusion features also improves PE-Core and SigLIP2 performance on GQA and LLaVA-Bench, indicating that these gains extend to general-purpose question answering.

We ablate concatenation vs. cross-attention at $s = 0$ (unconditional) and $s = 4$ (conditional) settings in Table 4. Simple concatenation with unconditional features already improves MMVP by 2.0 points, indicating that diffusion features encode complementary visual information. Increasing guidance to $s = 4$ yields further gains of +3.8% on MMVP and +3.4% on NaturalBench, demonstrating the added benefit of question-conditioned representations.

6. Conclusion

In this work, we investigate the potential of diffusion models to serve as task-aware feature extractors. We find that diffusion features encode rich vision-centric information that enables improved performance on vision-centric benchmarks. Next, we show that cross-attention maps capture fine-grained vision-text correspondences that propagate to block-level features, and that amplified guidance can leak

the prompt to the LLM. Building on this, we present Dif-fVLM, a simple yet effective fusion of CLIP and conditional diffusion features that leverages their complementary strengths and improves performance on both vision-centric and general-purpose benchmarks.

References

- [1] 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, Mikolaj 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
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen 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*, pages 382–398. Springer, 2016. 6
- [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] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, and Fabrizio Falchi. Is clip the main roadblock for fine-grained open-world perception?, 2024. 1
- [6] Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurate computation of the log-sum-exp and softmax functions, 2019. 5
- [7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025. 8
- [8] 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
- [9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 3
- [10] 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
- [11] 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. 5
- [12] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024. 2
- [13] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph 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), 2(3):6, 2023. 1
- [14] 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. 1, 3
- [15] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. 6
- [16] 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
- [17] 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, 4
- [18] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning, 2024. 1
- [19] 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
- [20] 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
- [21] Xuehai He, Weixi Feng, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Discdiffusion: Discriminative diffusion models as few-shot vision and language learners. *arXiv preprint arXiv:2305.10722*, 2023. 3, 5

- [22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2
- [23] 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
- [24] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6
- [25] 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. 1, 3
- [26] 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. 1
- [27] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024. 1
- [28] 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. 1
- [29] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 3
- [30] 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
- [31] Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4659–4669, 2025. 1, 3
- [32] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 6, 2
- [33] 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
- [34] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva 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. 4, 7, 2, 3
- [35] 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, 3
- [36] 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
- [37] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024. 1, 3
- [38] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing, 2024. 2
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3, 4, 1
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 4, 1
- [41] 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
- [42] 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
- [43] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prism: A vision-language model with multi-task experts. *arXiv preprint arXiv:2303.02506*, 2023. 1, 3
- [44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1
- [45] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi 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
- [46] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features, 2024. 5
- [47] 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
- [48] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024. 1

- [49] 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
- [50] 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
- [51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 4
- [52] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023. 1
- [53] 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
- [54] 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. 1, 3, 6
- [55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2
- [56] 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. 3, 6
- [57] 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
- [58] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. 6
- [59] 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
- [60] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention, 2022. 2
- [61] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality. In *CVPR*, 2022. 5
- [62] 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, 3
- [63] 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, 3, 4, 5, 6, 2
- [64] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 8
- [65] 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
- [66] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali 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*, pages 1921–1930, 2023. 2
- [67] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6
- [68] 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
- [69] 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. 6
- [70] 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
- [71] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 3

- [72] 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](#)
- [73] Runpeng Yu, Weihao Yu, and Xinchao Wang. Api: Attention prompting on image for large vision-language models, 2024. [1](#)
- [74] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. [1](#)
- [75] 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](#)
- [76] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. [6](#)
- [77] 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](#)

DiffVLM: Towards Text-Aware Multimodal Understanding via Stable Diffusion

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 5.

Table 5. **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

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 [39, 40]. The full hyperparameters are described in Table 6. 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 [13] model.

Table 6. **Training Hyperparameters.** Format from [63].

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 [44]	
DeepSpeed stage	2	2

Table 7. End-to-end latency comparison of vision encoders with LLaVA-v1.5-7B on a single A100 GPU.

Model	Vision Latency (ms)	LLM Latency (ms)	Total Latency (ms)
CLIP	47.96	314.48	362.44
SD	98.86	306.96	405.82
CLIP & SD Cross-Attn	155.66	316.59	472.25

B. Further Experimental Results

B.1. Computational Cost of Diffusion Model

To provide a quantitative assessment of the efficiency trade-offs, we benchmarked the end-to-end latency of our proposed model against the LLaVA-v1.5-7B baseline on a single A100 GPU.

The results as shown in Table 7 show that while our proposed fused vision encoder adds approximately 108 ms of latency compared to the standard CLIP encoder, this results in only a $\sim 30\%$ increase in the total end-to-end latency. This is because the LLM’s forward pass remains the primary computational bottleneck. Regarding scalability, we hypothesize that this vision encoder overhead will have a diminishing relative impact as language models continue to scale in size. Future research could explore more efficient fusion strategies for use cases requiring lower latency.

B.2. Comparison with Vision-Centric Models

We also compare diffusion features against other vision-centric models. We use DINOv2 as our baseline and show the results in Table 8. We find that SD2.1 and DINOv2 have similar performance on MMVP, but DINOv2 outperforms SD2.1 slightly on LLaVA-B and NaturalBench. On BLINK-val, SD2.1 obtains a 2.6% improvement. We show the BLINK-val breakdown in Table 9. Our findings show that diffusion features generally outperform DINOv2, with +7% on Multi-View, +9% on Reflectance, and +9% on Se-

semantic Correspondence.

B.3. Fusion Ablations

In Table 10, we explore the impact of using other visual backbones to improve upon CLIP. Specifically, we experiment with the DINOv2 and SigLIP2 vision encoders using the cross-attention fusion strategy. We observe that fusing conditional diffusion features outperforms other features across all tasks.

C. Full Captioning Results

We tabulate all COCO-Caption results displayed in Fig. 6 and present them in Table 11.

D. Further Diffusion Feature Analysis

D.1. Visualizations

We provide more visualizations of unconditional diffusion features via the NaturalBench benchmark [34] as it consists of more complex image-pairs compared to MMVP [63]. 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 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.

D.2. CKA Analysis

Here, we aim to understand the relationship between representations across different blocks. Specifically, we use CKA [32] 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 conditioning 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.

E. Further Cross-Attention Map Analysis

E.1. Quantifying Effect of Timesteps

We examine how different time-steps impact performance for MMVP-VLM in Table 12 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 13. 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.

E.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 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.

E.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, one set of maps is not necessarily better than

Table 8. Comparison of diffusion features with DINOv2 and CLIP across multiple benchmarks.

Config	Backbone	#Tokens	LLaVA-B	MMVP	BLINK-Val	NaturalBench
CLIP-LLaVA-v1.5-7B	ViT-L14/336	576	66.5	24.7	36.60	14.32
DINO-LLaVA-v1.5-7B	ViT-L14/224	256	48.9	22.7	34.66	9.95
SD-LLaVA-v1.5-7B	U-L1-R1-B0-Cross-Q, T=50	256	45.3	22.7	37.28	9.11

Table 9. BLINK-val task-wise results. Diffusion features outperform DINOv2 on Multi-View, Reflectance, and Semantic Correspondence.

Model	Backbone	Sim.	Count.	Depth	Jigsaw	Art	Func. Corr.	Sem. Corr.	Spat. Rel.	Obj. Loc.	Vis. Corr.	Multi -View	Refl.	Foren.	IQ	Avg.
CLIP-LLaVA-v1.5-7B	ViT-L/14-336	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	36.60
DINO-LLaVA-v1.5-7B	ViT-L/14-224	47.41	43.33	50.00	2.67	32.48	22.31	20.86	66.43	54.92	28.49	48.87	29.10	20.45	18.00	34.67
SD-LLaVA-v1.5-7B	U-L1-R1-B0, T=50	47.41	43.33	51.61	0.00	42.74	24.62	30.22	61.54	56.56	29.65	55.64	38.81	20.45	19.33	37.28

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.

F. 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-

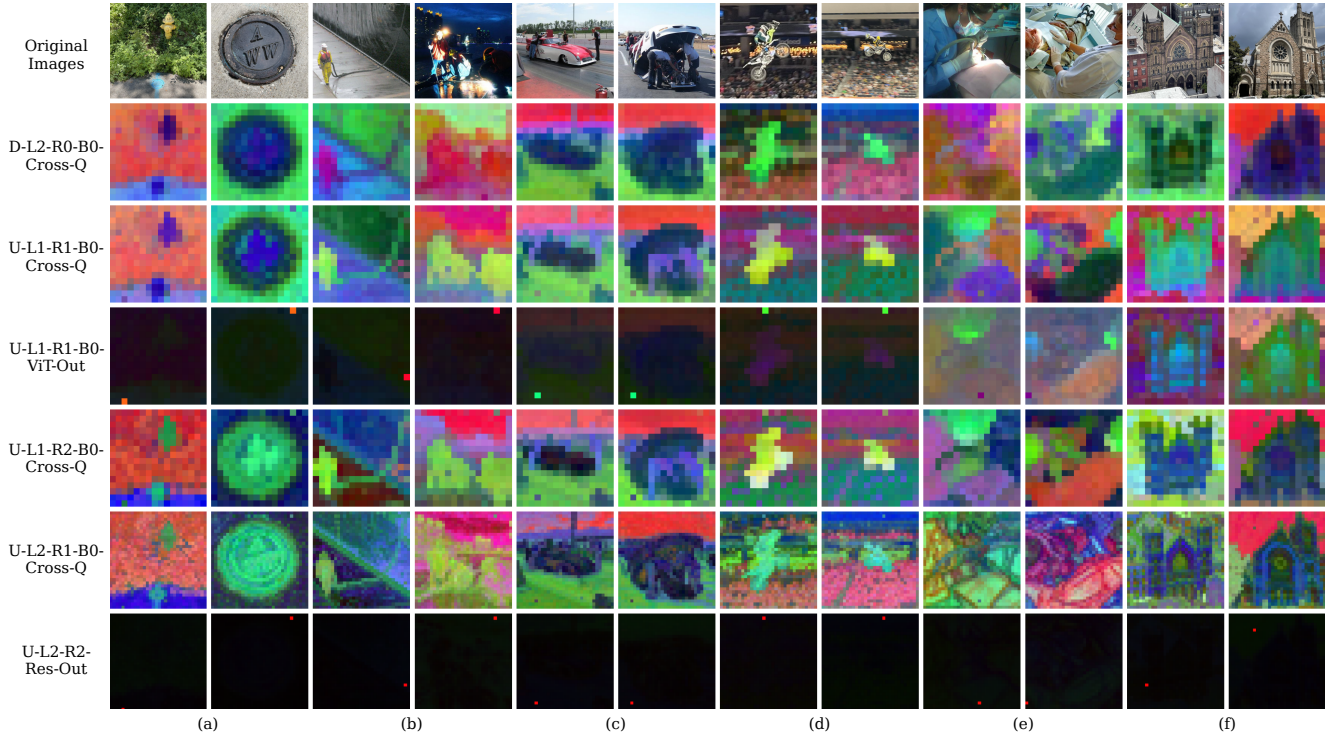


Figure 8. **More Diffusion Feature Visualizations.** We sample six pairs of images from the NaturalBench benchmark [34] and view the joint PCA maps across different blocks and layers. Please zoom in to see more details.

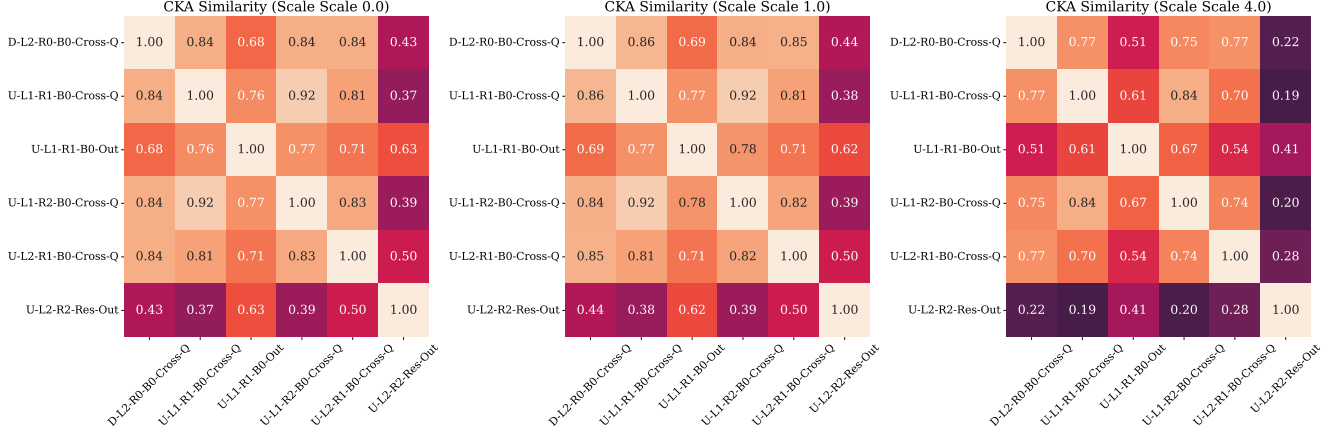


Figure 9. **CKA Block-Wise Similarity.** We compute block-wise CKA similarity using the COCO-Captions test set (5000 images) across various guidance scales ($s = 0, 1, 4$).

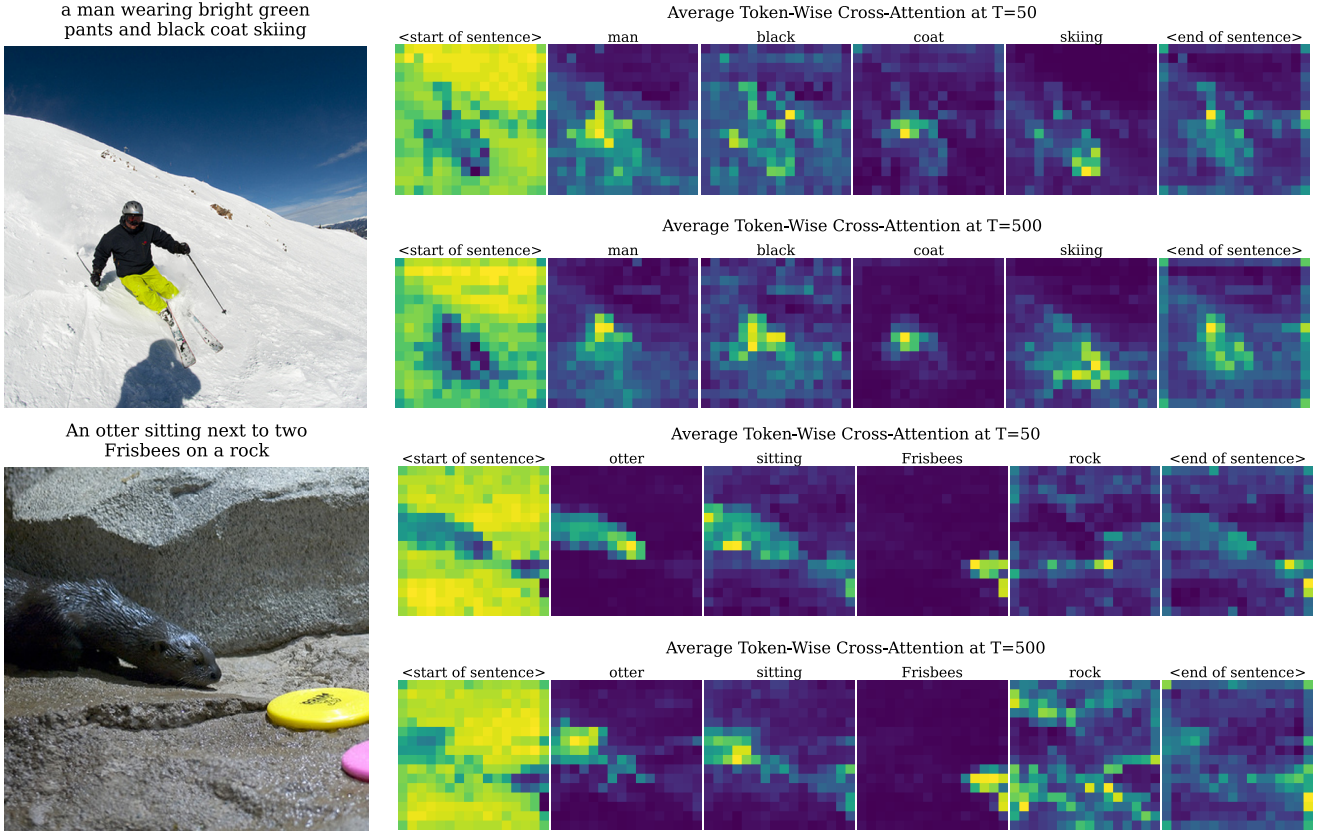


Figure 10. **More Cross-Attention Maps.** We display more examples of cross-attention maps from images in the COCO-Captions test set. For the image of the skier, we can see that object-attribute binding such as “black” and “coat” are better aligned at later timesteps compared to earlier ones.

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 fea-

tures 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

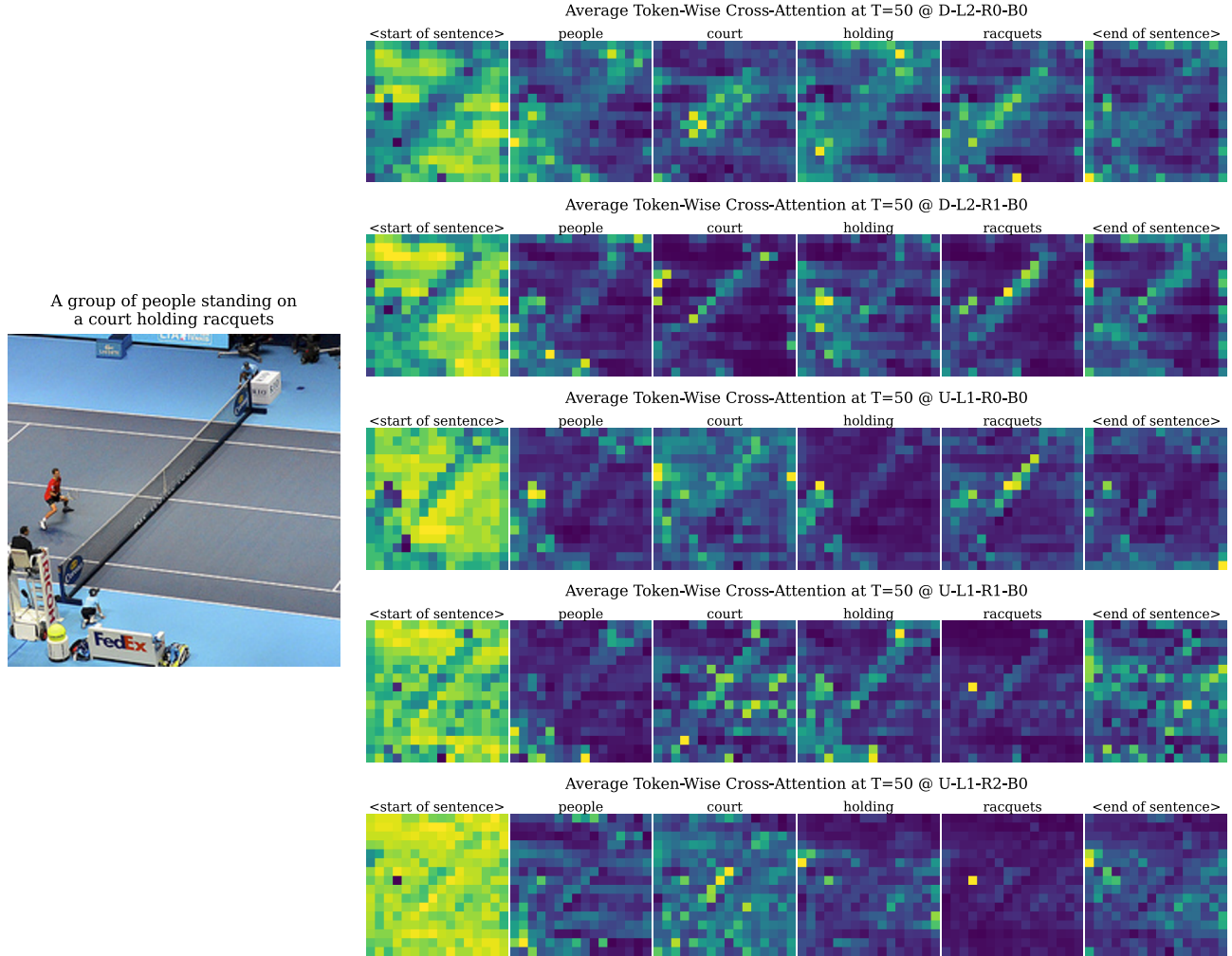


Figure 11. **Cross-Attention Maps Across Layers.** We display cross-attention maps at timestep 50 across various layers. We see that cross-attention maps are not uniform and that maps at the up-stage encode more robust image-text alignment.

Table 10. **Fusion Performance with other Visual Backbones** We compare performance when fusing CLIP features with conditional diffusion features as opposed to features from other strong vision encoders.

Config	MMVP	GQA	Natural-B
	Acc	Acc	G-Acc
OpenAI CLIP ViT-L-14	24.7	62.7	14.3
+ DINOv2 Cross-Attn	23.3	62.6	14.7
+ SigLIP2 Cross-Attn	23.3	58.6	11.9
+ SD (s = 4) Cross-Attn	30.7	62.5	16.2

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.

G. Investigating SDXL Architecture

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 [46]). For this analysis, only the projection layer is trained, while both the vision

Table 11. 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

Table 12. 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

encoder and LLM remain frozen. Our results are shown in Table 14. We observe that there are slight improvements in ROUGE and CIDEr scores when using SDXL, but SD2.1 outperforms SDXL on BLEU and SPICE.

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 15, 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.

H. Limitations

While our analysis demonstrates the potential of diffusion features for multimodal understanding, several limitations remain. First, our DiffVLM architecture introduces additional inference latency (30% overhead end-to-end), which may limit applicability in latency-sensitive settings. Second, despite our pre-training strategy where we provide no text, the question-conditioned diffusion features may still have some susceptibility to text leakage under high guidance scales. Future work could develop more effective strategies to handle this issue. Third, our experiments are conducted at moderate scale using LLaVA-1.5-7B with a fixed 256-token budget, and whether our findings generalize to larger MLLMs or higher token counts remains an open question.

Table 13. 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 $\pm$ 2.65	14.17 $\pm$ 1.15	10.50 $\pm$ 1.09

Table 14. 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

Table 15. Measuring 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

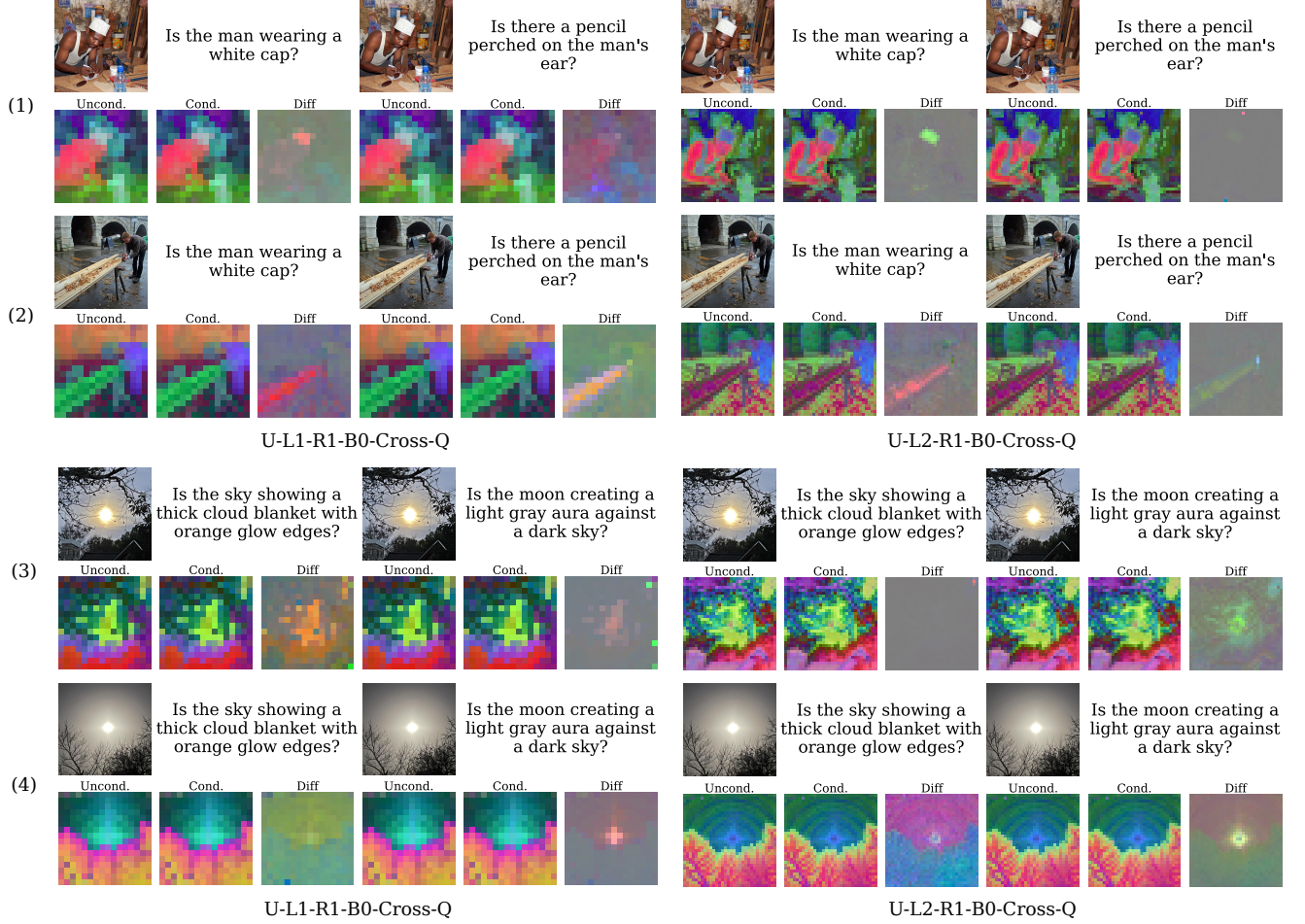


Figure 12. **More Visualizations of Question-Conditioned Features.** Please zoom in to see smaller regions of focus (e.g., pencil in second row, rightmost column of U-L2-R1-B0-Cross-Q)