Revelio: Interpreting and leveraging semantic information in diffusion models

Supplementary Material

A. Text Conditioning in Diffusion Models

Following the findings from [67], we report the performance of **Diff-C** in two text conditioning scenarios: i) empty prompt and ii) a meaningful prompt, e.g., "a photo of a $\{class\_name\}$, a type of pet", with the $class\_name$ first inferred through a zero-shot classification with CLIP. The motivation behind reporting both scores is to provide a comprehensive understanding of how text conditioning affects visual features at each layer. We report the classification performance with and without CLIP-inferred captions in Tables 6, 7, and 8. We note that passing specific class information inferred from CLIP generally helps across all three datasets, layers, and timesteps. To further understand how specific the captions should be, we experiment by passing a generic prompt, e.g., "a photo of a pet" during the diffusion process. As shown in Table 9 for `up_ft1` layer, on Oxford-IIIT Pet [44], compared to the base setting of passing in an empty prompt, using a generic prompt leads to a performance drop by 3.14%. This indicates that the specificity of the text being used to condition directly impacts feature representation quality, where more targeted prompts align better with class-relevant features, thereby improving model accuracy. Consequently, using precise text conditioning can lead to considerable gains in performance, particularly in distinguishing nuanced categories. However, this may not always be the case as described in Sec. 4.7, where for FGVC-Aircraft [38] conditioning with the class names led to a dip in classification performance.

B. Layer-wise PCA Analysis of Feature Maps

Figures 11 and 12 provides more evidence to the findings in Sec. 4.5. by highlighting differences in how SD 1.5 and DiT encode spatial information. In SD 1.5, the feature maps reveal well-defined spatial structures, with consistent colors and textures that correspond to specific regions in the image. By contrast, the feature maps of DiT display blended patterns, suggesting a stronger focus on capturing global context rather than emphasizing distinct spatial details.

C. Additional Details on the Visual Reasoning Task

Hyper-parameters: We adopt the same hyperparameters used in the the LLaVA-Lightning [34] configuration across all experiments. We use MPT-7B-Chat [58] as the language model, and CLIP ViT-L/14 [49], DINOv2 ViT-L/14 [43], SD 1.5 as the vision encoders. We show the training hyperparameters in Table 12. All experiments were conducted using a maximum of 4 NVIDIA RTX A6000 GPUs.

Pre-training datasets: Following LLaVA-Lightning [34], we use CC595k [55] for stage 1 pre-training, to align the visual encoder with the language model to establish a shared vision-language representation, by tuning the adapter. For stage 2 fine-tuning we use LLaVA-Instruct-80K [34] to fine-tune the model to enhance instruction-following capabilities.

Adapter settings: For experiments involving CLIP and DINOv2 features, we use the standard 2 layer MLP projector to align visual tokens with language tokens [34]. To obtain tokenized representations from the feature maps obtained from SD 1.5, we first add a 2 layer convolutional block and transform the feature map into pseudo-tokenized representations that match the token embedding dimensions of CLIP and DINOv2. These pseudo-tokenized representations are then passed into the 2 layer MLP projector for alignment.

Interleaving diffusion features with CLIP for visual reasoning tasks: For the experiments reported in Sec. 4.6, we first gradually reduce the spatial dimension of `up_ft1` from $1280 \times 32 \times 32$ to 256×1024 to match the token dimensions of CLIP vision embeddings. Next, we process these embeddings through two separate multi-layer projection layers resulting in projected embeddings of shape 256×4096. Finally, we interleave the projected token embeddings as done in [62] before passing them into LLaVA [34].

Performance: Table 11 compares the performance of different vision encoders in LLaVA, including CLIP (Table 11a), CLIP+DINOv2 (Table 11b), and CLIP+Diffusion at timesteps $t = 25$ (Table 11c) and $t = 200$ (Table 11d). The evaluation is conducted on the LLaVA-Bench (in-the-wild) [34] benchmark. The benchmark evaluates models across four categories: overall performance ('all'), complex reasoning ('LLaVA Bench complex'), conversational tasks ('LLaVA Bench conversational'), and descriptive tasks ('LLaVA Bench detail').

For the 'detail' category, CLIP+Diffusion at $t = 25$ achieves the highest relative score of $\mathbf{56.2}$, outperforming both CLIP (50.4) and CLIP+DINOv2 (37.7). This demonstrates that the interleaved diffusion and CLIP features effectively capture fine-grained visual details. In the

Timestep (t)	bottleneck (empty / from_CLIP)	up_ft0 (empty / from_CLIP)	up_ft1 (empty / from_CLIP)	up_ft2 (empty / from_CLIP)
0	52.27 / 52.74	48.79 / 49.06	64.09 / 62.826	50.55 / 49.15
25	51.76 / 54.88	50.49 / 51.19	65.07 / 63.69	50.25 / 49.21
100	51.07 / 55.12	49.48 / 51.10	64.15 / 64.98	51.37 / 50.53
200	50.91 / 52.51	49.63 / 49.99	63.88 / 63.13	50.53 / 50.55

Table 6. **Top-1 accuracy** at different timesteps and layers for fine-grained task (FGVC-Aircraft).

Timestep (t)	bottleneck (empty / from_CLIP)	up_ft0 (empty / from_CLIP)	up_ft1 (empty / from_CLIP)	up_ft2 (empty / from_CLIP)
0	68.79 / 84.33	66.99 / 79.50	88.28 / 90.11	77.79 / 80.67
25	69.97 / 85.25	73.29 / 84.17	88.61 / 90.68	81.63 / 85.28
100	69.53 / 85.88	67.07 / 81.33	88.29 / 90.97	78.82 / 84.36
200	68.03 / 86.43	65.87 / 81.79	86.89 / 90.32	78.49 / 84.79

Table 7. **Top-1 accuracy** at different timesteps and layers for fine-grained task (Oxford-IIIT Pet).

Timestep (t)	bottleneck (empty / from_CLIP)	up_ft0 (empty / from_CLIP)	up_ft1 (empty / from_CLIP)	up_ft2 (empty / from_CLIP)
0	85.83 / 92.13	85.75 / 89.68	86.75 / 88.18	79.11 / 80.28
25	87.59 / 91.32	86.54 / 90.81	87.72 / 91.08	81.07 / 82.69
100	88.28 / 92.41	88.18 / 90.66	87.99 / 92.05	82.02 / 84.73
200	88.36 / 91.65	87.23 / 90.73	89.22 / 92.26	82.69 / 85.63

Table 8. **Top-1 accuracy** at different timesteps for coarse-grained task (Caltech-101).

Prompt Type	Timestep	
	25	200
Empty Prompt	88.61	86.89
from_CLIP	2.34 ↑	3.94 ↑
generic	3.14 ↓	3.13 ↓

Table 9. **Performance vs Text Conditioning on Oxford-IIIT Pet using `up_ft1`:** Using a generic prompt ("A photo of a pet") leads to a dip in classification performance compared to using an empty prompt. By contrast, using a targeted caption ("A photo of a $\{class\_name\}$, a type of pet") leads to a boost in performance.

Layer	Output Shape	Description
conv1	$[B, 1024, H, W]$	conv2D, 3×3
conv2	$[B, 1024, H/2, W/2]$	conv2D, 3×3, stride 2
conv3	$[B, 1024, H/4, W/4]$	conv2D, 3×3, stride 2
conv4	$[B, 1024, H/8, W/8]$	conv2D, 3×3, stride 2
GAP	$[B, 1024, 1, 1]$	global average pooling
FC	$[B,$ NUM_CLASSES$]$	flatten + FC layer

Table 10. Architecture of **Diff-C** ($40M$ params).

'complex' category, CLIP+Diffusion at $t = 200$ achieves the highest relative score of $\mathbf{70.5}$, surpassing CLIP (68.4). At $t = 25$, CLIP+Diffusion scores 67.9 indicating that the coarser-grained features extracted at higher timesteps ($t = 200$) seem more effective for this specific task that requires broader contextual understanding. Next, for the 'conversational' category, CLIP+Diffusion at $t = 25$ achieves a relative score of $\mathbf{51.3}$, outperforming both CLIP (43.9) and CLIP+DINOv2 (35.6). The interleaving of diffusion and CLIP features significantly enhances the model's ability to handle visually grounded conversational tasks effectively.

Finally, we report the overall performance under the 'all' category and note that CLIP+Diffusion achieves a superior performance with a score of $\mathbf{59.9}$ at $t = 25$, outperforming CLIP's standalone score of 56.6. This reinforces the power of the visual representations learnt from the diffusion process in achieving top-performance on diverse vision-language tasks.

D. Additional k-SAE Visualizations

DiT vs U-Net: In this section, we provide additional visualizations of k-SAE features. As shown in Fig. 13 (b), (e), `Block 14` of DiT captures more class-specific information than other blocks which is qualitatively corroborated in Table 2d. However, compared to SD 1.5, DiT captures less distinct class information, as seen in the snow background in Fig. 13 (h). Moreover, the spatially related photographic styles observed in Sec. 4.5 do not emerge in DiT. We hypothesize that the transformer-based relies less on inductive bias information compared to UNet-based SD 1.5, as discussed in Sec. 4.5.

Later timesteps: Figure 14 presents k-SAE visualization at $t = 500$ for SD 1.5. Compared to $t = 25$, features at $t = 500$ focus more on low-level information, such as texture and low-light, which is qualitatively corroborated in

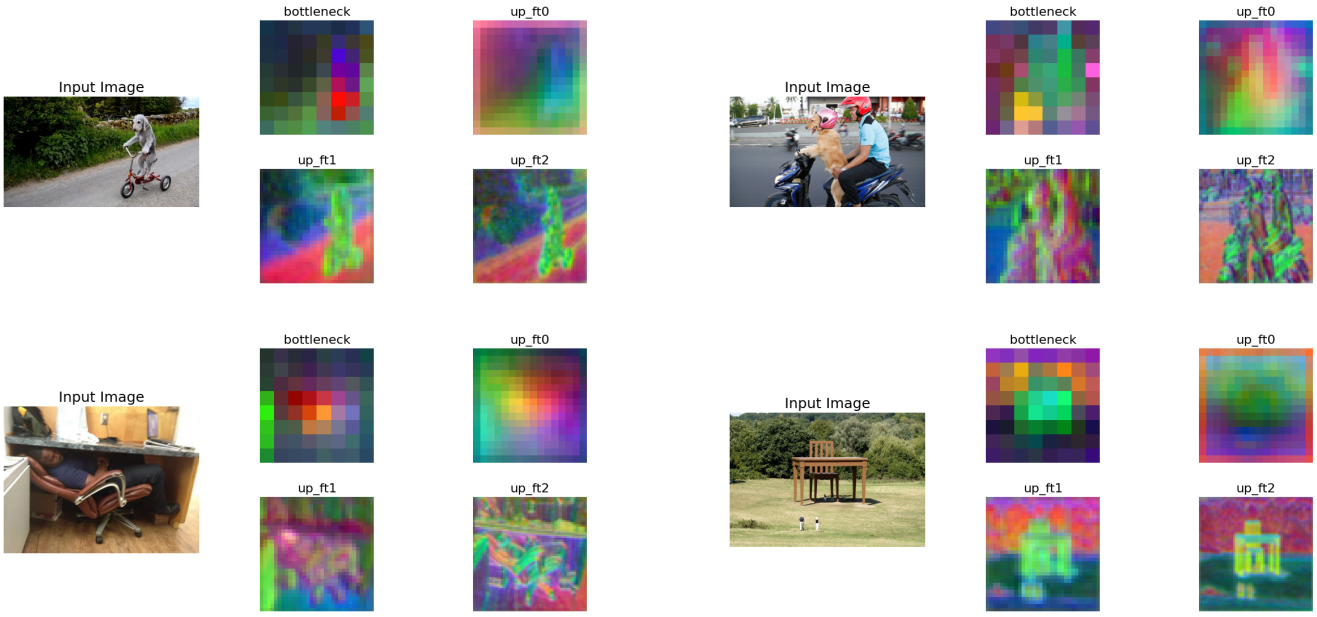

Figure 11. **PCA Feature Maps SD-1.5 on images** from UnRel [47] - Consistency of colors and textures (at up_ft1, up_ft2) suggests that the model preserves local details and spatial relationships

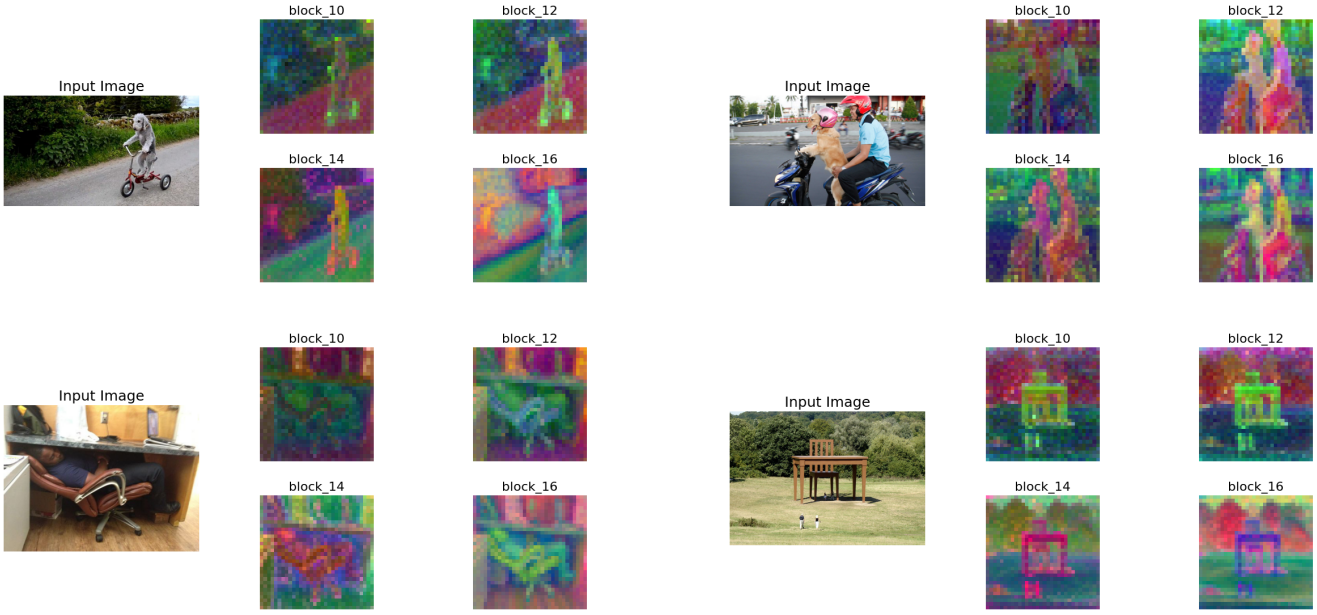

Figure 12. **PCA Feature Maps DiT on images from UnRe**l [47] - The blending of colors suggests that the model encodes global relationships while maintaining a holistic representation of spatial structures, rather than isolating precise local details.

Table 2b. We hypothesize that as the diffusion timestep increases, so does the added noise, rendering the features less useful for transfer learning, consistent with our observations in Sec. 4.4.

E. Additional k-SAE Experiments

Effect of image resolutions: To assess the robustness of our method across varying input resolutions, we conduct additional experiments examining the effect of image resolution. As shown in Table 13, using different image resolutions exhibits a similar trend in terms of σ_{label}, with smaller resolutions resulting in slightly reduced variance across dif-

Category	Relative Score	GPT-4 Score	LLaVA Score
All	56.6	82.7	46.8
LLaVA Bench complex	68.4	80.4	55.0
LLaVA Bench conversational	43.9	87.1	38.2
LLaVA Bench detail	50.4	82.0	41.3

(a) CLIP LLaVA

Category	Relative Score	GPT-4 Score	LLaVA Score
All	47.0	84.8	39.8
LLaVA Bench complex	59.9	81.1	48.6
LLaVA Bench conversational	35.6	94.1	33.5
LLaVA Bench detail	37.7	81.3	30.7

(b) CLIP+DINOv2 LLaVA

Category	Relative Score	GPT-4 Score	LLaVA Score
All	**59.9**	83.2	49.8
LLaVA Bench complex	67.9	80.0	54.3
LLaVA Bench conversational	**51.3**	90.6	46.5
LLaVA Bench detail	**56.2**	80.7	45.3

(c) CLIP+Diffusion ($t = 25$) LLaVA

Category	Relative Score	GPT-4 Score	LLaVA Score
All	56.8	83.7	47.5
LLaVA Bench complex	**70.5**	80.0	56.4
LLaVA Bench conversational	45.6	87.7	40.0
LLaVA Bench detail	45.7	86.0	39.3

(d) CLIP+Diffusion ($t = 200$) LLaVA

Table 11. **Performance on the multi-modal reasoning task for various LLaVA configurations**. The integration of Diffusion features with CLIP improves performance across all tasks, with notable gains in the 'detail' and 'conversational' categories.

Hyperparameter	Stage	
	Stage 1	Stage 2
batch size	128	128
learning rate (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 [35]	
deepspeed stage	2	3

Table 12. Hyperparameters for LLaVA-Lightning (default setting)

ferent DiT blocks on Oxford-IIIT Pet.

F. Additional Implementation Details

In this section, we provide additional implementation details for training k-SAE. We set the expansion factor for the k-SAE to 64, following prior work [22], resulting in $n = 1280 \times 64 = 81,920$ latents for SD and $n = 1152 \times 64 = 73,728$ latents for DiT. We apply a unit normalization constraint [54] on the decoder weights W_{dec} of the k-SAE after each update. We use the Adam [31] optimizer with a learn-

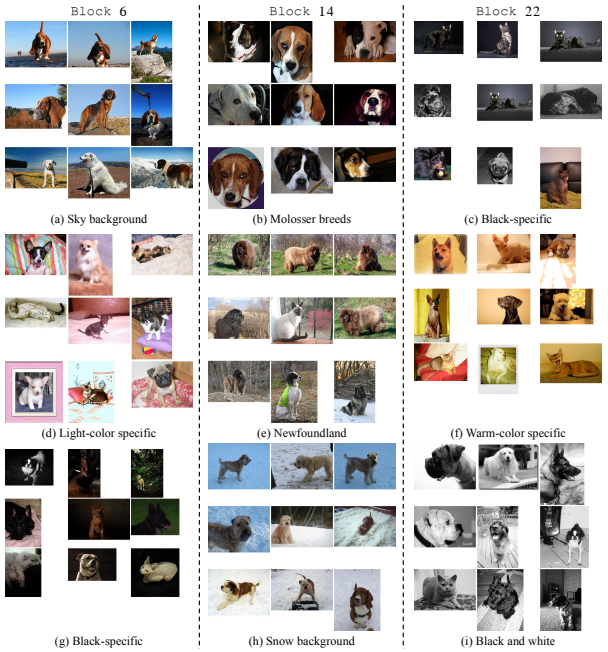

Figure 13. **k-SAE visualizations of the blocks on Oxford-IIIT Pet** at $t = 25$. `Block 14` mainly captures class-specific information, while other blocks focus more on less distinct features.

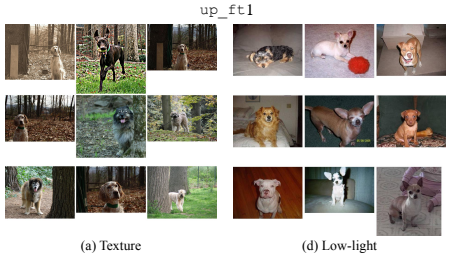

Figure 14. **k-SAE visualizations on Oxford-IIIT Pet** of `up_ft1` UNet layer at $t = 500$. In contrast to the earlier timestep (Fig 3), $t = 500$ appears to focus more on low-level features.

Block	Oxford-IIIT Pet	
	(512)	(256)
6	10.18	10.36
10	9.44	10.16
14	**9.05**	**10.06**
18	9.55	10.11
22	9.84	10.13

Table 13. **Label purity** (σ_{label}) measured by computing the average standard deviation of the class labels of the top-10 most highly activating images among the top 1000 most highly activating features of the learned k-SAEs for different DiT blocks with different resolutions on Oxford-IIIT Pet. Lower is better.

ing rate of 0.0004 and apply a constant warm up for 500 steps. The total training time is approximately 1 hour with $\sim$18 GB peak memory on 1 NVIDIA RTX A6000 GPU trained for $10M$ steps.

G. Additional Details of Evaluation

In this section, we provide additional details on how we quantify the granularity of semantic information in diffusion features through a multiple-choice question-answering task, as discussed in Sec. 4.3. Using GPT-4o [3], we evaluate the level of semantic detail captured by different diffusion features. Table 14 presents the prompt used to query the model for this evaluation. Specifically, we assess the model's predictions based on the top 10 most highly activating images among the top 100 most highly activating neurons of the learned k-SAE.

```
Prompt: Each set of images captures
    different types of patterns:
1. Class-specific information (e.g., fine-
    grained details, animals of the same
    breed).
2. Moderately granular features (e.g.,
    similar-looking animals irrespective
    of their position).
3. Very coarse information (e.g.,
    foreground objects similarly placed
    relative to the background).
4. Could not detect patterns (e.g., noisy
    or no specific patterns).
Select only one number (1, 2, 3, or 4)
    that best describes the shared pattern
**Respond with just the number and nothing
     else.**
```

Table 14. Input prompt for GPT-4o based evaluation.