Bridging Generative and Discriminative Models for Unified Visual Perception with Diffusion Priors

Shiyin Dong[1] , **Mingrui Zhu**[1,*] , **Kun Cheng**[1] , **Nannan Wang**[1,*] , **Xinbo Gao**[2]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
shiyindong@stu.xidian.edu.cn, mrzhu@xidian.edu.cn, kunncheng@stu.xidian.edu.cn, nnwang@xidian.edu.cn, gaoxb@cqupt.edu.cn

Abstract

The remarkable prowess of diffusion models in image generation has spurred efforts to extend their application beyond generative tasks. However, a persistent challenge exists in lacking a unified approach to apply diffusion models to visual perception tasks with diverse semantic granularity requirements. Our purpose is to establish a unified visual perception framework, capitalizing on the potential synergies between generative and discriminative models. In this paper, we propose Vermouth[1], a simple yet effective framework comprising a pre-trained Stable Diffusion (SD) model containing rich generative priors, a unified head (*U-head*) capable of integrating hierarchical representations, and an *Adapted-Expert* providing discriminative priors. Comprehensive investigations unveil potential characteristics of Vermouth, such as varying granularity of perception concealed in latent variables at distinct time steps and various U-net stages. We emphasize that there is no necessity for incorporating a heavyweight or intricate decoder to transform diffusion models into potent representation learners. Extensive comparative evaluations against tailored discriminative models showcase the efficacy of our approach on zero-shot sketch-based image retrieval (ZS-SBIR), few-shot classification, and open-vocabulary (OV) semantic segmentation tasks. The promising results demonstrate the potential of diffusion models as formidable learners, establishing their significance in furnishing informative and robust visual representations.

1 Introduction

Over the years, there has been a keen interest in the representation learning capabilities of generative models, given their proficiency in generating vivid images. Consistent with the perspective that a generative model must attain the semantic understanding ability to produce high-fidelity samples. Early works [Vincent *et al.*, 2008; Vincent *et al.*, 2010] have demonstrated that generative models can be employed for discriminative tasks through non-trivial methods.

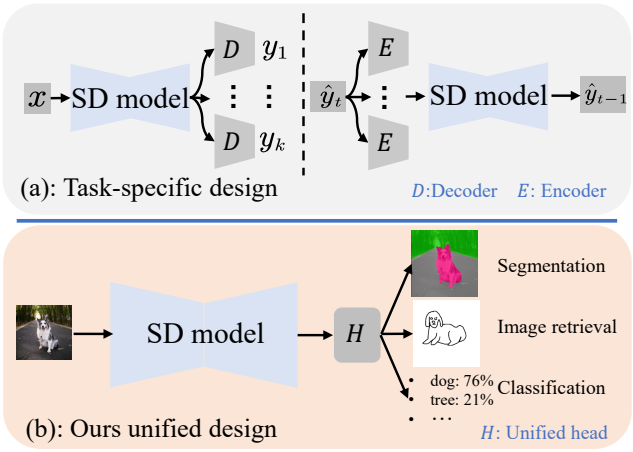


Figure 1: We transfer priors of the SD model in a unified framework for different visual perception tasks.

Recently, diffusion models have emerged with stunning performance in the image generation field, creating realistic and incredibly detailed images [Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Ho *et al.*, 2020]. Specifically, large-scale text-to-image diffusion models [Rombach *et al.*, 2022; Parmar *et al.*, 2023] can seamlessly integrate and modify semantic information in an end-to-end manner, which allows the synthesis of images featuring diverse objects, scenes, and styles [Zhang *et al.*, 2023b; Brooks *et al.*, 2023]. Given this phenomenon, we deem that large text-to-image diffusion models, such as SD [Rombach *et al.*, 2022] model, have acquired high-level and low-level semantic cues through extensive exposure to large-scale image-text pairs. Nevertheless, the question of how to extract the latent knowledge embedded in the diffusion process and harness this knowledge for visual perception tasks remains an unsolved challenge.

Visual perception tasks necessitate the establishment of distinct decision boundaries $p_\theta(y|x)$ among categories, an objective not initially envisioned in the design of diffusion

*Corresponding author

[1]Our model is named after vermouth. Indeed, the properties of diffusion and blending of wines and herbs in vermouth share similarities with our design.

models. This achievement is typically attained through supervised learning [Liu *et al.*, 2021; Liu *et al.*, 2022], unsupervised contrastive learning [Caron *et al.*, 2021], and masked image modeling followed by supervised fine-tuning [He *et al.*, 2022; Bao *et al.*, 2021]. In contrast, diffusion models aspire to model the inherent probability distributions $p_\theta(x|z)$ of a dataset. Consequently, an inherent incompatibility exists between diffusion models and visual perception tasks. As shown in Figure 1, various methods [Zhao *et al.*, 2023; Karazija *et al.*, 2023] have attempted to address this issue by integrating off-the-shelf and heavy decoders or encoders with SD models. Nevertheless, the exploration of how to effectively leverage hierarchical features within SD models in a unified framework are inadequate.

The goal of this paper is to scrutinize the release of internal priors within diffusion models and their transfer to non-generative tasks in a unified manner. In contrast to approaches involving various decoders, we propose a simple yet effective method applicable to tasks with diverse granularity requirements within a unified framework. Specifically, for the input image, we introduce suitable noise and project it into the latent space of the SD model under accurate text guidance. Subsequently, ours U-head blends latent representations from different granularities, eliminating the need for designing complex and tailored decoders. Moreover, this architecture demonstrates remarkable flexibility, enabling smooth integration with the priors of Adapted-Expert, resulting in enhanced compatibility with visual perception tasks. Leveraging the flexibility and effectiveness of this module and the rich semantic features embedded within diffusion, we can seamlessly transfer the fused features to diverse tasks.

To comprehensively analyze the most effective way for unlocking the knowledge within the SD model, we investigated three downstream tasks across a range of near-real-world scenes, including ZS-SBIR, OV semantic segmentation, and few-shot classification tasks to evaluate our method. Experiment results on over 20+ datasets reveal that, despite its inherent mismatch with visual perception tasks, the SD model can still be regarded as a promising learner.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to propose a unified framework to apply diffusion to visual perception tasks demanding different granularity semantics.
- We design a unified head capable of effectively fusing the generative priors of SD models with discriminative priors from the Adapted-Expert.
- Comprehensive experiments and analyses unveil observed rules for hyperparameters such as the noise level of latents, which can provide constructive insights and suggestions for future researches.

2 Related Work

2.1 Vision Models in Perception Task

The paradigm of pre-training and transfer learning has significantly advanced the domain of computer vision. In the early years, convolutional neural networks [He *et al.*, 2016] and Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] were viewed as standard architectures for various vision tasks, due to their exceptional perception capabilities. Nevertheless, the conventional pre-training on fixed-label datasets constrains their applicability in some complex scenarios.

Recently, CLIP [Radford *et al.*, 2021] has garnered significant attention in various fields as its efficient visual language alignment pre-training. Some approaches leverage CLIP for tasks such as few-shot classification [Gao *et al.*, 2023], image retrieval [Saito *et al.*, 2023], and image segmentation [Zhou *et al.*, 2022], which demonstrates the remarkable and rapid learning capabilities. In this paper, our focus shifts to the SD model, aiming to investigate whether it has the same ability and how to harness this potential capabilities.

2.2 Perceptual Learning with Diffusion Models

The research of transferring generative models to discriminative tasks, exemplified by BigBiGAN [Donahue and Simonyan, 2019], has sustained substantial interest over an extended period. Diffusion models [Ho *et al.*, 2020], such as the SD model, have not only excelled in the field of image synthesis but have also drawn considerable attention in other fields. Harnessing latent features, the SD model exhibits versatile applications in diverse domains, including classification [Wei *et al.*, 2023], image segmentation [Xu *et al.*, 2023; Li *et al.*, 2023], and depth estimation [Ji *et al.*, 2023; Zhao *et al.*, 2023]. However, these methods require task-specific design such as encoders and decoders that makes them complex. In contrast, we sidestep this tailored procedure and advocate a unified framework capable of accommodating various scenarios.

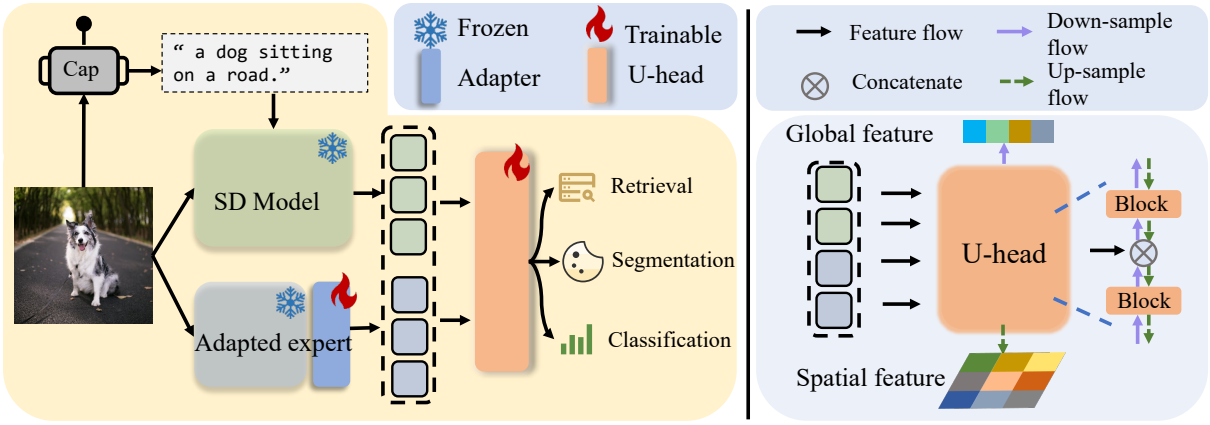
We noted that the methods most akin to ours are VPD [Zhao *et al.*, 2023] and Grounded-Diffusion [Li *et al.*, 2023]. VPD employs tailored decoders for distinct tasks and utilizes ambiguous prompts for text guidance. Grounded-Diffusion relies on an additional pre-trained grounding model and employs iterative denoising from pure Gaussian noise to get clean images. In comparison to VPD, our architecture is more lightweight, unified, and flexible. Compared with Grounded-Diffusion, our application scenarios are broader and do not necessitate additional model-assisted.

3 Method

To directly apply the SD model in non-generative domains within a unified framework, we propose Vermouth. It is devised to transfer the prior knowledge of the SD model for different visual perception tasks, eliminating the need for task-specific designs. We will start with the preliminaries of diffusion models in Section 3.1. Subsequently, the U-head for obtaining the final representation and the Adapted-Expert to enhance compatibility between the SD priors and discriminative tasks are detailed in Section 3.3 and Section 3.4.

3.1 Preliminaries

Diffusion models constitute a category of likelihood-based models non-equilibased on brium thermodynamics. These models can characterize the data distribution $p(x)$ by learning to reverse the forward process, which incrementally introduces noise to the data. The forward diffusion process at



(a) Overall framework (b) U-head details

Figure 2: An overview of our framework. We employ the pre-trained BLIP model to acquire an accurate description and utilize the SD model to generate representations guided by text embedding. The introduction of the U-head is intended to fuse the two representations from the SD model and the Adapted-Expert, aiming to enhance compatibility in discriminative tasks with different semantic granularity requirements.

t-th time-step can be modeled as Markov: $z_t \sim q(z_t|z_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}z_{t-1}, (\beta_t)\boldsymbol{I})$, where β_t is associated with the noise schedule [Ho *et al.*, 2020]. By employing a reparameterization trick, we can simplify this expression into a more manageable form:

$$q\left(z_t \mid z_0\right) = \mathcal{N}\left(z_t; \sqrt{\bar{\alpha}_t} z_0, \left(1-\bar{\alpha}_t\right) \boldsymbol{I}\right), \tag{1}$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s = \prod_{s=1}^{t} (1-\beta_s). \tag{2}$$

Generally, diffusion models introduce noise to inputs until $z_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and samples iteratively by denoising the latent variables:

$$p(z_{0:T}) = p(z_T) \prod_{t=1}^{T} p(z_{t-1}|z_t). \tag{3}$$

Through the proper simplification, one can predict the noise component ϵ by neural network $\epsilon_\theta(x_t; t)$ implemented by a U-net to learn how to reconstruct the input data:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{z_t, \epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(z_t; t, c)\|_2^2 \right], \tag{4}$$

where c is an additional condition such as text prompts.

The SD model employs a VQ-VAE [Van Den Oord *et al.*, 2017] encoder to project the input into the latent space, denoted as $z_0 = \mathcal{E}(x)$. Subsequently, a decoder is utilized to reconstruct the input, expressed as $\hat{x} = \mathcal{D}(\hat{z_0})$. For latent modeling, the SD model adopts an asymmetric U-net, which is structured into three stages: down-sample, bottleneck, and up-sample, encompassing a total of 18 blocks. See the supplementary for the architectural details of the SD model.

3.2 Latent Prior in SD Model

Given the presence of multi-scale pattern in the U-net, one can extract desired features in specific blocks.

$$F_{\text{SD}} = \text{U-net}(z_t, t, c), \tag{5}$$

where $F_{\text{SD}} = \{f_i \in \mathbb{R}^{H_i \times W_i \times C_i} | i \sim \boldsymbol{B}\}$ and $\boldsymbol{B}$ is indexes of the specific blocks.

Recent works [Zhao *et al.*, 2023; Xu *et al.*, 2023] have opted for different blocks, combining with specific decoders to extract features for discriminative tasks. However, there remains a lack of clear insight regarding which blocks' output are more semantically rich. This challenge stems from the non-trivial nature of selecting specific blocks among a total of 18. We opt for a macro level to investigate the semantics within the SD model by viewing the stages as the research granularity to explore semantic information at various levels.

$$f_i = \text{U-net}_{\texttt{stage\_i}}(z_t, t, c), \tag{6}$$

where `stage_i` represent the stages index.

As a text-to-image model, the text prompt plays a crucial role in feature extraction as it serves as guidance for semantic synthesis. An intuitive approach involves using all class names s in the dataset $\mathcal{D}$ to form the text context:

$$c = \mathcal{T}(\text{concat}([s|s \in \mathcal{D}])), \tag{7}$$

where $\mathcal{T}$ is the text encoder of CLIP. However, employing the global context may lead to potential misalignment, especially considering that different inputs will contain different objects. In contrast, we enhance alignment by utilizing BLIP [Li *et al.*, 2022] caption model as shown in Figure 2 to derive the image-aligned text prompt $s = \text{Cap}(x)$. The multi-modally pre-trained BLIP model, for a given image, excels in generating accurate descriptions while preserving semantics, encompassing both the object and its surroundings present in the image $s \in \mathcal{X} \subset \mathcal{D}$.

In addition to the output features f_i at each level, the attention maps of U-net may also capture semantics. Inspired by recent work [Zhao *et al.*, 2023] considering cross-attention map as the diffusion prior equally, we are motivated to investigate the impact of the cross-attention map $A = \text{Softmax}(\frac{QK^T}{\sqrt{d}})$, where Q and K denote the image and text hidden states. However, we find that this approach did

not consistently yield positive results, as detailed in our experimental findings.

3.3 U-Head for Perception Tasks

Our U-head is designed to receive multi-scale features extracted from the SD model, facilitating the capture of language-aware visual features. As illustrated in Figure 2, to obtain global features, this module progressively fuses high-resolution features containing detailed semantics to low-resolution features with global semantics along the down-sample flow:

$$h = \text{U-head}(F), \tag{8}$$

where h is the final representation. Contrarily, detailed pixel-level features can be acquired along the up-sample flow. This approach encourages the capture of visual features at various levels of granularity, seamlessly blending coarser to finer semantic cues. Consequently, the need for customized redundant decoders[2] is mitigated to a certain extent. Then, the final output can be obtained through attention pooling or a single convolution layer noted by $v = W \cdot h$.

We follow several works [Gu *et al.*, 2021; Gao *et al.*, 2023] that utilize the output features of the CLIP text encoder as the final classifier weight to align v with text features $t = \mathcal{T}(S)$, where S means prompt constructed by categories names. This strategy is beneficial to the creation of a more favorable feature space for recognition [Radford *et al.*, 2021]. We ensure the alignment through cosine similarity:

$$p(y|x) = \frac{v \cdot t}{\|v\| \cdot \|t\|}. \tag{9}$$

This method offers flexibility for expanding to unseen labels by merely adjusting the input prompt S and facilitates the learning of language-aware visual representations.

3.4 Combining the Discriminative Prior

To effectively transfer learned features to discriminative tasks while ensuring compatibility, an intuitive approach is to introduce the prior knowledge of the recognition model. Leveraging the high flexibility of our U-head, we can readily fuse the diffusion prior with the discriminative prior:

$$h = \text{U-head}([F_{SD}; F_{exp}]), \tag{10}$$

where $[\cdot;\cdot]$ and F_{exp} means concatenation and discriminative prior respectively.

We employ ResNet-18 [He *et al.*, 2016] to introduce discriminative prior $F_{exp} = \text{ResNet}(x)$ due to its inherent possession of multi-resolution features. Thanks to the flexibility of our method, in fact, any discriminative model such as DINO-v2 [Oquab *et al.*, 2023] can be introduced. More results can be found in supplementary material.

Furthermore, to enhance the integration between discriminative and generative prior, we incorporated an Adapter [Houlsby *et al.*, 2019] following the discriminative model as illustrated in Figure 2. Experiments demonstrate improved performance with the adapted features. Recognizing its role in enhancing compatibility with visual perception tasks, we refer to it as the "Adapted-Expert".

[2]We use the terms decoder and head to distinguish modules of different magnitudes. In general, the decoder is larger than the head and has more parameters.

3.5 Details of Training

Initially, for the input image, we employ BLIP to obtain an accurate description. Subsequently, we employ the text encoder to derive the conditions, denoted as$c = \mathcal{T}(s)$. Following the method proposed in the previous sections, the final representations are obtained as Equation 10. After getting the prediction according to Equation 9, the cross-entropy loss is applied for training:

$$\mathcal{L} = \frac{1}{N} \sum^{N} \hat{y} \log p(y|x). \tag{11}$$

4 Experiments

We verify the effectiveness of our method compared to other traditional vision models over 20+ different datasets grouped into 3 tasks, including ZS-SBIR, OV semantic segmentation, and few-shot classification. In addition, we also show our results on a faster training schedule and validate the effectiveness of each key component of our method.

4.1 Experimental Settings

Unless specified, we use SD 1-5 [Rombach *et al.*, 2022] and freeze all the parameters of the SD model to preserve latent knowledge. For system-level comparisons, we select a few typical methods that employ different pre-training strategies, such as DINO [Caron *et al.*, 2021] (contrastive learning), ConvNeXt [Liu *et al.*, 2022], Swin-Transformer [Liu *et al.*, 2021] (supervised learning), MAE [He *et al.*, 2022] (masked image modeling), and BeiTv3 [Wang *et al.*, 2023] (Multi-modality learning).

ZS-SBIR. Following the general setting [Liu *et al.*, 2019], we report the mean Average Precision (mAP) of our method on three datasets: Sketchy, TU-Berlin, and QuikDraw. Given the heterogeneity of the sketch and image domain, this task serves as a robustness test. We train our model for 1 epoch, and the learning rate is set to 1e-4.

OV Semantic Segmentation. Following the general setting [Zhou *et al.*, 2022], we train on the COCO-Stuff dataset and evaluate on the validation set of five datasets: ADE20K-150 (ADE-150), ADE20K-847 (ADE-847), Pascal VOC (VOC), Pascal Context-59 (PC-59), and Pascal Context-459 (PC-459). Training our model for 120k iterations under weak data augmentation default and 8k iterations for the fast schedule, we report the mean Intersection over Union (m-IoU) at a single scale.

Few-shot Classification. Following the setting of CLIP-Adapter [Gao *et al.*, 2023], we report the 16-shot classification accuracy on 11 datasets: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, EuroSAT, UCF101, DTD, and SUN397. We train our model for 100 epochs by default and 10 epochs for a fast training schedule.

Detailed information on training procedures and datasets can be seen in the supplementary material.

4.2 Main Results

In this section, we present the main results and compare with the counterparts under the default settings. The configuration

config	Classification	Segmentation	ZS-SBIR
prompt	BLIP		
time steps	200	10	200
attention map	w. up cross-att. map		w.o cross-att. map
clip proj	w.o. projection		w. projection
noise schedule	ddpm schedule		ddim inv
stage in U-net	mid + down		mid + up

Table 1: Main configuration of our method.

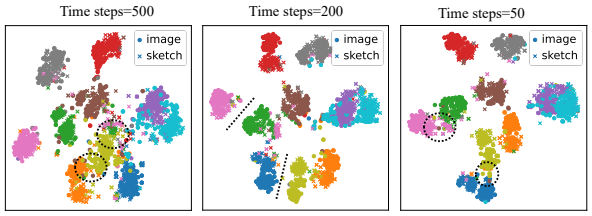


Figure 3: Feature visualization of different categories of sketches and images in different time steps.

across three tasks is outlined in Table 1. Each row corresponds to a key configuration. Detailed explanations of each configuration provided in Section 4.3.

ZS-SBIR

ZS-SBIR is an appealing task as it resembles real-world scenarios. For a given sketch query x_s, natural images x_i of the same category are retrieved based on feature similarity, which is usually measured by mAP. This task requires models to balance domain heterogeneity between sketch and image and identify unseen categories. To fairly assess the performance, we augment the counterpart models with an Adapter and align visual features with the text features, as mentioned in Equation 9 to facilitate knowledge transfer and mitigate the semantic gap.

As demonstrated in Table 2, our method outperforms all the traditional models, indicating that internal features in the SD model are more adept at handling abstract representations of sketches. This phenomenon is illustrated in Figure 3, where the disparity between images and sketches diminishes as a result of light perturbation introduced at the appropriate time step (e.g., $t = 200$). This leads to a more regular distribution within the feature space.

OV Semantic Segmentation

OV semantic segmentation is the most challenging task, requiring the model trained on limited categories to achieve precise recognition of arbitrary or even noisy categories (e.g., ADE847 contains 847 categories, most of which are noisy).

For the counterpart models, we employ UperNet [Xiao *et al.*, 2018] architecture and fine-tune the entire model by default. As shown in Table 3, we achieve comparable or even superior results compared to the tailored methods that combine specific decoder and discriminant backbone. It's noteworthy that we only have 5.9 million learnable parameters and therefore do not possess an advantage in terms of the learnable parameters. This is attributed to the fact that we only fine-tuned the "`Normalization`" layer (only 0.2 million parameters) in U-net. However, leveraging the U-head with efficient fusion capabilities and the excellent representation of the SD model, we still achieved a better performance against some tailored methods.

model \ mAP	Sketchy	TU-Berlin	QuickDraw
MAE-L	39.23	41.99	11.71
BeiTv3-G	54.54	50.93	13.67
Swinv2-L	43.39	45.51	12.08
DINO-B	38.51	25.49	10.15
Vermouth	**56.8**	**52.83**	**15.11**

Table 2: Main results on ZS-SBIR task. Compared to traditional visual models, we achieve the best results, which is marked in **bold**

model \ m-IoU	# param	ADE-150	PC-59	VOC20	ADE-847	PC-459	COCO
MAE-L	442	17.5	53.27	93.51	3.42	8.82	44.88
ConvNeXt-L	235.3	18.65	**53.42**	94.62	3.53	**9.53**	48.0
Swin-L	234	18.8	53.37	**94.76**	**3.8**	9.42	49.41
DINO-B	144.4	17.13	47.84	92.44	3.16	7.75	42.78
Vermouth	**5.9**	**19.0**	52.88	92.87	3.7	9.0	**46.44**

Table 3: OV semantic segmentation results on six datasets, where param means learnable parameters. We achieve comparable results while minimizing the number of tunable parameters. Results on the training set are marked in gray

.

Few-Shot Classification

Few-shot classification is a challenging task that demands learning from only a few samples and generalizing efficiently across multiple scenarios. For the competitor model, we fit a linear layer (i.e., linear prob) to assess its generalization ability. Due to space constraints, we only report the top-1 accuracy in the 16-shot setting.

As shown in Table 4, our method outperforms MAE on all datasets except for UCF101, OxfordPets, and SUN397. Notably, in comparison to the MAE pre-trained on IN-21K, we achieve 16.15% improvement on IN-1K. While the overall performance does not exceed the discriminative models, comparable results are attained on specific datasets, such as FGVCAircraft and EuroSAT. This indicates that the inherent advantage of discriminative models in recognition tasks, due to paradigm differences with generative models, is diminishing without using any additional tricks.

An interesting observation reveals that SD models generally do not perform well on fine-grained datasets since the model may not fully understand the distinction of each class at a fine-grained level. Some failure cases are illustrated in the supplementary material.

4.3 Sensitive Analysis

In this section, we explore the potential factors affecting the performance of our method through ablation studies conducted on a faster schedule in ImageNet, Sketchy, and ADE20K (semantic segmentation). Table 5 reveals several intriguing properties.

model \ acc@1	OxfordPets	Flowers102	FGVCAircraft	DTD	EuroSAT	StanfordCars	Food101	SUN397	Caltech101	UF101	ImageNet	Avg
MAE-L	91.87	92.04	36.51	63.74	87.39	24.15	59.31	62.08	94.45	76.55	39.74	66.17
BeiTv3-G	**93.79**	97.84	38.34	72.41	86.11	**62.58**	74.42	71.57	96.9	**84.38**	**86.95**	**78.66**
Swinv2-L	89.65	**99.61**	29.13	**73.1**	86.9	37.75	**77.41**	**72.63**	**97.01**	81.06	78.84	74.83
DINO-B	89.32	97.82	**48.3**	69	**91.15**	57.17	58.5	62.44	95.57	76.97	67.66	73.99
Vermouth	66.13	92.35	42.52	66.62	88.93	51.05	45.78	58.09	95.83	70.49	55.89	66.74

Table 4: Main results of 16-shot learning on 11 datasets. Compared to MAE, we achieve better results on some datasets and diminish the gap with professional discriminative models.

stage in U-net	IN-1K	Sketchy	ADE20K
down	36.82	49.8	33.34
up	40.12	54.02	35.39
down + mid	**42.73**	55.76	36.88
up + mid	42.52	**56.8**	**41.28**
all	42.04	53.66	40.53

(a) **Stage in U-net**. Combining mid stage with the up-sample or down-sample stage brings better results.

prompt	IN-1K	Sketchy	ADE20K
null	29.46	44.71	39.23
random	30.04	52.28	40.55
BLIP	**42.74**	**56.8**	**41.28**

(b) **Prompt**. An image-aligned prompt provides better guidance.

clip proj	IN-1K	Sketchy	ADE20K
w.o. proj	**42.74**	56.41	**41.28**
w. proj	37.62	**56.8**	40.07

(c) **CLIP projection**. The text features after projected work better.

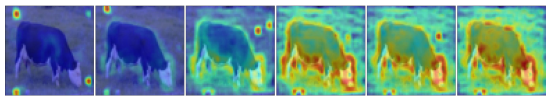
noise	IN-1K	Sketchy	ADE20K
w.o.	38.59	47.44	39.87
ddim inv	42.08	**56.8**	41.01
ddpm schedule	**42.74**	54.97	**41.28**

(d) **Inversion method**. The results of different inversion methods have slight differences.

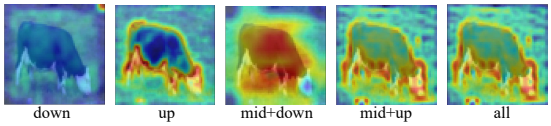
attention	IN-1K	Sketchy	ADE20K
w.o.	41.32	**56.8**	41.03
down	42.07	56.5	40.41
up	**42.74**	56.57	**41.28**
up + down	42.08	54.51	40.09

(e) **Cross-attention map**. Cross-attention map doesn't always seem to bring benefits.

Table 5: Sensitive analysis of potential factors affecting model performance in three tasks. The default setting is marked in **bold**



(a) Time step



(b) Stage in U-net

Figure 4: Visualization of features captured by our model at different times steps and stages of U-net

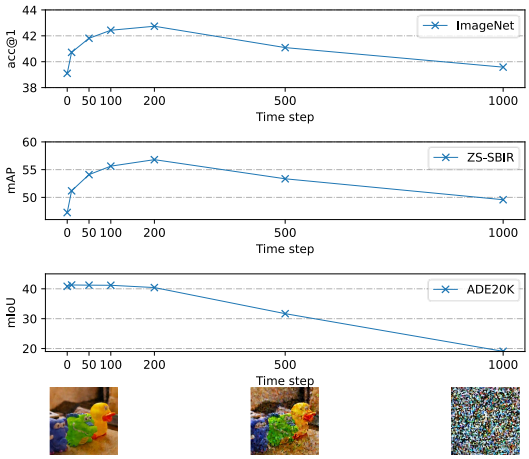


Figure 5: Different time steps bring different results for three tasks.

Time steps. To validate the semantic properties of the latents under different noise-level, we evaluate our model across a series of time steps. In Figure 5, we observe a consistent trend in classification and image retrieval tasks, indicating that latent representations at intermediate time steps $t \in [100, 200]$ yield better performance. However, this phenomenon is not presented in the segmentation task. This characterization is consistent with the visualization of different time steps in Figure 4a. Specifically, as the images exhibit slight corruption, leading to the removal of fine-grained details, intermediate time steps focus more on capturing subject-level information, whereas forward time steps exhibit a perception of the entire image. Therefore, for tasks like semantic segmentation, which demand for local semantics, superior results are achieved within the time steps $t \in [10, 100]$ as these steps preserve the detailed features.

Stage in U-net. As outlined in Section 3.1, the SD model comprises distinct stages, each encompassing different features. We selected individual stages or their associated combinations at stage granularity and examined their performance. As shown in Table 5a, exclusively extracting features from a single stage yields sub-optimal results due to

the incompleteness of semantics. However, there is a marked improvement when combining features from the mid-stage with either the up-sample or down-sample stage. This phenomenon is consistent with Figure 4b, illustrating that different stages encapsulate semantics at various granularities. Combining the mid stage with either the up-sample or down-sample stages results in more comprehensive semantic fusion and, consequently, superior performance.

Text prompt. The key component of our method is the image-aligned prompt. In comparison to an empty prompt (F_{SD} = U-net$(z_t, t, c = \varnothing)$), image-aligned prompts can accurately reflect the content of the image and provide precise guidance. As shown in Table 5b, the prompt obtained by BLIP brought 12.7%, 4.52%, and 0.73% boosts on the three tasks compared to the random prompt. We attribute this to the fact that the image-aligned prompt is consistent with the usage of pairwise data pre-training for SD models, which allows us to retain the semantic bootstrapping abilities, resulting in better performance reasonably. Contrarily, an empty or random text prompt will lead to potentially harmful guidance.

CLIP token. Some methods [Zhang *et al.*, 2023a] have demonstrated that using the output of the second-last layer in ViT performs better. In the case of the CLIP, the final layer is the `projection` layer. We are interested in exploring whether a similar phenomenon occurs in our model, given the significant role of the CLIP text encoder in our method. As illustrated in Table 5c, utilizing the token from the second-to-last layer (wo. proj) leads to enhancements of 5.12% and 1.21% in classification and segmentation respectively. This implies that unprojected text features exhibit clearer decision boundaries and are more suitable as final classifier weights. However, performance on Sketchy has a slight decrease, which we hypothesize to be potentially associated with the diverse perceptual characteristics of different layers in the CLIP text encoder [Gandelsman *et al.*, 2023].

Inversion method. Recently, certain approaches [Mokady *et al.*, 2023] have revealed that distinct inversion strategies for acquiring noisy inputs result in varied outcomes in image generation tasks. Therefore, we juxtapose two inversion methods, specifically, DDIM Inversion [Song *et al.*, 2020], and the DDPM schedule [Ho *et al.*, 2020]. DDIM inversion inverts latent variable z_0 into its corresponding noised version z_t by continuousizing ordinary differential equations (ODEs). DDPM schedule adds randomness to the latent variables under the control of β_t.

As illustrated in Table 5d, the disparities between the DDPM schedule and DDIM inversion are minimal. This observation suggests that the discriminative cues within the SD model remain largely unaffected by the method of noise incorporation. However, employing the SD model to extract features from a clean image is suboptimal. This inefficiency is attributed to the fact that clean latent variables do not belong to the latent space of the SD model as it is trained to predict clean images from noised version.

Cross-attention map. Given the mechanism of injecting semantics guidance through the cross-attention in the SD model, we were prompted to investigate whether such a mechanism could also be advantageous in recognition scenarios. We select the average cross-attention map from different stages in the U-net and concatenate it with our image feature, denoted as $F = [F_{SD}; F_{exp}; A]$.

	ZS-SBIR		Segmentation	16-shot Classification	
	Sketchy	Avg	ADE20K*	IN-1K	Avg
baseline	54.28	40.29	40.3	37.49	57.32
+fuse	55.43	40.99	40.88	42.38	59.86
+expert	**56.8**	**41.44**	**41.28**	**55.89**	**66.74**

Table 6: An ablation study on the key components validates the effectiveness of our approach. * means trained under the fast schedule

As outlined in Table 5e, only a subtle differences in performance have been observed. The cross-attention map located in the up-sample stage yields an improvement of 1.42% and 0.25% in IN-1K and ADE20k but registers a decrease of 0.23% in Sketchy. This may be caused by inaccurate attention due to the abstract nature of the sketch.

Discussion. We analyzed to identify the potential factors influencing the behavior of our model. The experimental results indicate the existence of shared optimal settings applicable to both dense prediction and global recognition tasks, such as text prompts. However, certain factors, including the cross-attention and clip projection, result in inconsistent responses.

4.4 Ablation Study

We conducted ablation studies on our two pivotal designs, U-head and Adapted-expert. To assess the effectiveness of U-head, we established a baseline using a straightforward fusion technique, involving a convolution operation on the obtained diffusion features followed by a global summation. A consistent average improvement of $0.7\%, 0.58\%$, and 2.54% is observed when compared to the baseline, which indicates that U-head ensures both structural unity and effectiveness. Furthermore, when combined with the Adapted-Expert, an additional improvement of $0.45\%, 0.4\%$, and 6.88% is observed across all three tasks. This indicates that our method, following knowledge fusion with the discriminative model, acquires more discriminative cues, thereby enhancing discriminative accuracy. In summary, experiments across all tasks verify the effectiveness of the crucial design in our method.

5 Conclusion

We introduce Vermouth, a simple yet effective unified framework that is designed to transfer the generative priors of diffusion models to discriminative tasks. Leveraging the BLIP model, we capture description as context conditions, preserving the inherent advantage of semantic guidance of SD models. To accommodate various downstream tasks, we introduce a lightweight head capable of seamlessly integrating discriminative and diffuse representations within a unified framework. Experiments involving multi-tasks conducted on the unified architecture illustrate the generality and efficiency. Through the careful selection of time steps and other key components, Vermouth effectively migrates the rich visual semantics of the SD mode in downstream classification, retrieval, and segmentation tasks and demonstrates a promising performance. This exploration will not only offer valuable guidance on harnessing and optimizing the potential of SD models but also inspire further research into developing more efficient frameworks.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2023YFA1008600, in part by the National Natural Science Foundation of China under Grant 62106184, Grant U22A2096, and Grant 62036007; in part by the Young Talent Fund of Association for Science and Technology, Shaanxi, China, under Grant 20230121; in part by Xi'an Science and Technology Plan Project under Grant 23GJSY0004; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042 and Grant KYFZ24012.

References

[Bao *et al.*, 2021] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Donahue and Simonyan, 2019] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gandelsman *et al.*, 2023] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023.

[Gao *et al.*, 2023] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

[Gu *et al.*, 2021] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[Ji *et al.*, 2023] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. *arXiv preprint arXiv:2303.17559*, 2023.

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

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023.

[Liu *et al.*, 2019] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3662–3671, 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Parmar *et al.*, 2023] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[Radford *et al.*, 2021] 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.

[Rombach *et al.*, 2022] 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.

[Saito *et al.*, 2023] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[Wang *et al.*, 2023] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186, June 2023.

[Wei *et al.*, 2023] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*, 2023.

[Xiao *et al.*, 2018] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[Xu *et al.*, 2023] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[Zhang *et al.*, 2023a] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023.

[Zhang *et al.*, 2023b] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhao *et al.*, 2023] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.

[Zhou *et al.*, 2022] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.