Diff-Feat: Diffusion-Based Cross-Modal Feature Extraction for Multi-Label Classification

Tian Lan[1], Yiming Zheng[1], Jianxin Yin[1,†]
[1] *Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China*
[†] *Corresponding author: jyin@ruc.edu.cn*

***Abstract*—Multi-label classification has broad applications and depends on powerful representations capable of capturing multiple interactions. We introduce *Diff-Feat*, a simple yet effective framework that extracts intermediate features from pre-trained diffusion-Transformer models for both images and text, and fuses them for downstream tasks. We observe distinct patterns in optimal feature selection: for vision tasks, the most discriminative intermediate feature along the diffusion process occurs at the middle step and is located at the middle block in Transformer. In contrast, for language tasks, the best feature occurs at the noise-free step and is located at the deepest block. Notably, a consistent phenomenon across diverse datasets: a fixed "Middle Layer" yields the best performance on various downstream classification tasks for images (under DiT-XL/2-256×256 and various U-ViT architectures). We further analyze the underlying mechanism and its scaling behavior across model scales. To efficiently select representations, we devise a heuristic local-search algorithm that pinpoints the locally optimal "image-text"×"block-timestep" pair among a few candidates, avoiding an exhaustive grid search. A simple fusion-linear projection followed by addition-of the selected representations achieves 46.1% mAP on Visual Genome 500 (with 500 categories) and 80.2% mAP on Pix2Cap-COCO (with 80 categories), consistently outperforming strong baseline methods. Code is available at https://github.com/lt-0123/Diff-Feat.**

***Index Terms*—Diffusion models, multi-label classification, multimodal fusion, representation learning, Transformer**

I. Introduction

Multi-label classification has wide-ranging applications in image retrieval, biomedical image recognition [1], and scene understanding [2]. Compared with single-label classification, it poses two major challenges [3]: label imbalance and the difficulty of extracting features from regions of interest, which demands more expressive feature representations.

Recent advances in diffusion models [4, 5, 6] have demonstrated remarkable generative capabilities across multi-modal domains such as image synthesis [7, 8], audio generation [9, 10], and natural language processing [11, 12]. Beyond generative modeling, diffusion models have recently attracted growing interest as representation learners for downstream tasks [13, 14, 15], leveraging their denoising process to learn rich semantic features [13]. The likelihood-based training objective gives the model strong potential to function as a foundation model in a self-supervised setting.

Furthermore, for language modeling, diffusion-based language models offer greater flexibility than conventional autoregressive models in capturing diverse language-order dependencies by supporting any-order generation [16], thereby opening new possibilities for learning more comprehensive and robust semantic representations. This property also makes diffusion language models a promising foundation for extracting high-quality representations for downstream analysis and understanding tasks.

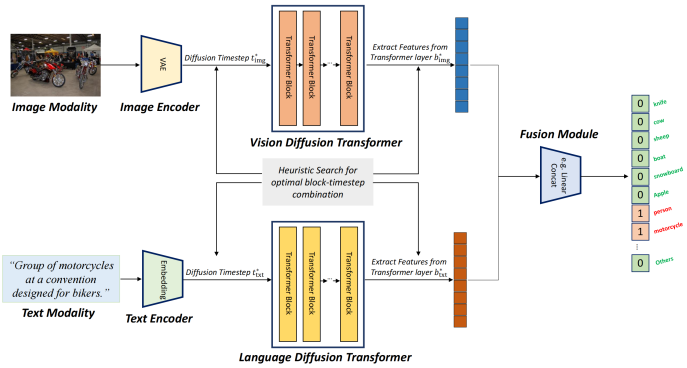


Fig. 1: Overview of the Diff-Feat framework.

Inspired by the strong linear separability and semantic understanding capabilities of diffusion-based representations [14], a central question arises: *How can we identify the optimal pair of diffusion-based representations from image and text to enhance the performance of downstream multi-label classification?* To address this question, we propose *Diff-Feat*, a simple yet effective framework for multi-label classification. As shown in Fig. 1, Diff-Feat extracts visual and textual features from pre-trained continuous diffusion-Transformer models across different noise levels and Transformer blocks, and then fuses them to perform multi-label prediction.

To the best of our knowledge, we are the first to treat image and text modalities symmetrically by independently extracting their diffusion representations. We conduct a comprehensive empirical study that characterizes the effectiveness of representations extracted at different noise levels and Transformer blocks across both modalities. Furthermore, we propose a simple heuristic-guided local search algorithm to efficiently identify the optimal image-text representation pair. When fused, the selected representations achieve leading performance on multi-label classification benchmarks: 46.1% mAP on Visual Genome 500 [17] and 80.2% mAP on Pix2Cap-COCO [18].

Meanwhile, we observe a consistent and robust phenomenon: for image data, regardless of the diffusion timestep,

dataset distribution, downstream classification task, evaluation metric, or model architecture and scale, the most discriminative features consistently emerge from the **middle** Transformer block of image diffusion-Transformer models. We further investigate the underlying mechanisms and examine how this pattern scales across various diffusion-Transformer architectures.

Our main contributions are summarized as follows:

- We propose *Diff-Feat*, a simple yet effective framework that extracts cross-modal diffusion representations for multi-label classification. Using a heuristic strategy to identify optimal fusion points, our method achieves strong performance on Visual Genome 500 and Pix2Cap-COCO.
- We present a large-scale and systematic empirical study that characterizes how Transformer depth and diffusion noise jointly shape representation quality across visual and textual modalities. Our analysis spans multiple datasets, diffusion timesteps, and probing tasks, providing principled insights into modality-specific robustness patterns and the emergence of discriminative semantics.
- We systematically uncover a robust and architecture-agnostic phenomenon: across diffusion timesteps, datasets, downstream tasks, evaluation metrics, model scales, and visual diffusion architectures, representations from the **middle** Transformer layers consistently yield the strongest discriminative performance. Through extensive scaling experiments and modality-specific mechanistic analyses, we show that this effect is tightly linked to model depth in vision models and driven by noise-signal trade-offs in vision diffusion systems.

II. Background and related work

Multi-label classification. Multi-label classification is a supervised learning task where an instance can be associated with multiple labels. In recent years, various deep learning techniques have been applied to address multi-label classification. Existing approaches explore diverse strategies for multi-label classification, including Transformer-based querying [3], graph-based modeling [19], and multimodal alignment [20]. Meanwhile, cross-modal approaches [21, 22, 23, 24, 25] improve multi-label classification performance while effectively alleviating overfitting on majority classes. However, effectively addressing imbalanced label distributions and capturing complex label dependencies remains challenging and largely unresolved problems.

Diffusion models and diffusion representations. Diffusion models [5] define a forward process in which an input $\mathbf{x}_0$ is progressively corrupted by Gaussian noise over a series of timesteps $t = 1, \cdots, T$. At each timestep, the noisy sample $\mathbf{x}_t$ is sampled from the distribution $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$, where $\alpha_t = \sqrt{\prod_{i=1}^{t}(1-\beta_i)}$ and $\alpha_t^2 + \sigma_t^2 = 1$. The noise schedule $\beta_1, \ldots, \beta_T$ is determined by a linear schedule from $\beta_{\min}$ to $\beta_{\max}$, i.e.,

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon} \tag{1}$$

Inspired by the representation power of Denoising Autoencoders (DAEs) [26, 27] in compressed latent spaces, recent work [13, 14, 28, 29] has increasingly explored the representation learning potential of diffusion models. However, existing research has primarily focused on images, with limited attention given to extracting representations from language-based diffusion models.

III. Approach

A. Discriminative diffusion representations for image and text

Inspired by prior work [14] that leverages intermediate activations from pre-trained image diffusion models, we adopt a similar philosophy. This strategy requires no modification to standard diffusion backbones and remains fully compatible with existing models.

Based on this idea, we also utilize intermediate activations extracted from pre-trained continuous language diffusion models [30, 31], focusing on specific decoder blocks and noise levels. Unlike previous approaches [15, 32] that treat text as a conditional embedding $\tau(\mathbf{s})$ and extract intermediate activations via $f = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \tau(\mathbf{s}), t)$, where $\tau(\mathbf{s})$ denotes the embedding of the image caption $\mathbf{s}$ by a pre-trained text encoder τ and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ denotes the diffusion model parameterized by $\boldsymbol{\theta}$, we treat both text and image as equal modalities for representation learning. This symmetric strategy provides greater flexibility for selecting task-relevant features from different layers and noise levels.

To apply noise, we randomly sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and apply Eq. 1 to obtain $\mathbf{x}_t$ for images, since no significant differences are observed between random and deterministic noising methods [14]. However, for text, we find that deterministic noising (e.g., DDIM [33]) yields better performance.

Formally, we define the problem as identifying the optimal diffusion timestep $t \in \mathcal{T}$ and decoder block $b \in \mathcal{B}$ that minimize the discriminative loss on a downstream task, i.e., $(t^*, b^*) = \arg\min_{t \in \mathcal{T}, b \in \mathcal{B}} \mathcal{L}(t, b)$, where $\mathcal{L}(t, b)$ denotes the downstream discriminative loss, $\mathcal{T}$ and $\mathcal{B}$ denote the sets of diffusion timesteps and decoder blocks, respectively.

We conduct linear probing to identify diffusion representations with strong linear separability and label semantics. In addition, we fine-tune the image diffusion model to enhance downstream performance.

B. Fusion strategy with unimodal diffusion representations

Furthermore, the performance of multi-label tasks can be significantly improved by selecting the optimal "image-text"×"block-timestep" pairs and employing an effective fusion method. However, performing a grid search over all possible block-timestep combinations is computationally impractical, with a high complexity of $O(|\mathcal{T}|^2|\mathcal{B}|^2)$, where $|\cdot|$ denotes the cardinality of the set.

To address this challenge, we adopt a heuristic search (Algorithm 1), where img and txt denote the image and text modalities, respectively. We first identify the optimal configuration for each modality, thereby reducing the search space by focusing on high-potential candidates. We then conduct

a localized grid search within the neighborhoods of these unimodal optima to find the best fusion configuration. This approach significantly lowers computational cost while maintaining competitive performance, with a reduced complexity of $O(|\mathcal{T}||\mathcal{B}|)$.

Algorithm 1 Heuristic Local Search for Fusion Block-Timestep Selection

Input: Candidate blocks $\mathcal{B}$, timesteps $\mathcal{T}$, evaluation functions `EvalImg`, `EvalTxt`, `EvalFus`
Output: Optimal fusion block-timestep pair $((b^*_{\text{img}}, t^*_{\text{img}}), (b^*_{\text{txt}}, t^*_{\text{txt}}))$

Step 1: Identify unimodal optima
$(b'_{\text{img}}, t'_{\text{img}}) \leftarrow \arg\max_{b_{\text{img}} \in \mathcal{B},\, t_{\text{img}} \in \mathcal{T}} \texttt{EvalImg}(b_{\text{img}}, t_{\text{img}})$
$(b'_{\text{txt}}, t'_{\text{txt}}) \leftarrow \arg\max_{b_{\text{txt}} \in \mathcal{B},\, t_{\text{txt}} \in \mathcal{T}} \texttt{EvalTxt}(b_{\text{txt}}, t_{\text{txt}})$

Step 2: Construct local neighborhood
$\mathcal{C} \leftarrow \text{Neighbors}\big((b'_{\text{img}}, t'_{\text{img}}), (b'_{\text{txt}}, t'_{\text{txt}})\big)$

Step 3: Local fusion search
$((b^*_{\text{img}}, t^*_{\text{img}}), (b^*_{\text{txt}}, t^*_{\text{txt}})) \leftarrow \arg\max_{((b_{\text{img}}, t_{\text{img}}), (b_{\text{txt}}, t_{\text{txt}})) \in \mathcal{C}} \texttt{EvalFus}(b_{\text{img}}, t_{\text{img}}, b_{\text{txt}}, t_{\text{txt}})$
return $((b^*_{\text{img}}, t^*_{\text{img}}), (b^*_{\text{txt}}, t^*_{\text{txt}}))$

Let $\mathbf{h}_{\text{img}} \in \mathbb{R}^{d_{\text{img}} \times 1}$ and $\mathbf{h}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}} \times 1}$ denote the diffusion representations from image and language continuous diffusion pre-trained models, respectively, where d_{img} and d_{txt} represent the dimensionalities of image and text features, which need not be equal. We explore several strategies to combine them before feeding them into the multi-label classifier: (1) directly concatenating them, i.e., Concat$(\mathbf{h}_{\text{img}}, \mathbf{h}_{\text{txt}})$; (2) firstly performing a linear projection to $\mathbf{h}_{\text{img}}$ and $\mathbf{h}_{\text{txt}}$, then concatenating them, i.e., Concat$(\mathbf{W}_{\text{img}}\mathbf{h}_{\text{img}}, \mathbf{W}_{\text{txt}}\mathbf{h}_{\text{txt}})$, where $\mathbf{W}_{\text{img}} \in \mathbb{R}^{d_{\text{alg}} \times d_{\text{img}}}$, $\mathbf{W}_{\text{txt}} \in \mathbb{R}^{d_{\text{alg}} \times d_{\text{txt}}}$, and d_{alg} denotes the dimensionality of the shared alignment space for image and text representations; (3)firstly performing a linear projection to $\mathbf{h}_{\text{img}}$ and $\mathbf{h}_{\text{txt}}$, then adding them, i.e., $\mathbf{W}_{\text{img}}\mathbf{h}_{\text{img}} + \mathbf{W}_{\text{txt}}\mathbf{h}_{\text{txt}}$; (4) Cross attention: the image representations $\mathbf{h}_{\text{img}}$ are used as queries, while the text features $\mathbf{h}_{\text{txt}}$ serve as both keys and values, i.e., CrossAttention$(\mathbf{h}_{\text{img}}, \mathbf{h}_{\text{txt}}) = \text{softmax}\Big(\frac{\mathbf{W}_Q\mathbf{h}_{\text{img}}(\mathbf{W}_K\mathbf{h}_{\text{txt}})^\top}{\sqrt{d_k}}\Big)\mathbf{W}_V\mathbf{h}_{\text{txt}}$, where $\mathbf{W}_Q \in \mathbb{R}^{d_k \times d_{\text{img}}}, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_k \times d_{\text{txt}}}$ are learned projection matrices, and d_k is the key dimensionality.

IV. Experiments

Datasets. We consider two primary multi-modal, multi-label datasets: VG500 [34], a widely used subset of Visual Genome [17], and Pix2Cap-COCO [18]. We aggregate the region-level captions within each image into a single textual description, which is used as the corresponding image-level caption in our experiments. We further evaluate our method on additional benchmark datasets, including MS-COCO (2014) [35], CIFAR-100 [36], Tiny-ImageNet [37], and PASCAL VOC 2007 [38], to assess its robustness across different dataset scales and label configurations.

Evaluation metrics. We evaluate multi-label classification performance using a comprehensive set of metrics, including mean average precision (mAP), per-class precision (CP), per-class Recall (CR), per-class F1 (CF1), overall precision (OP), overall recall (OR) and overall F1 (OF1).

Linear classifier setting. The extracted diffusion features are evaluated using a single-layer linear classifier trained with Binary Cross-Entropy (BCE) loss for multi-label tasks or Cross-Entropy loss for single-label tasks. For the linear probing setting, the classifier is trained using the Adam optimizer with an initial learning rate of 1×10^{-3}, together with a cosine annealing learning rate schedule over 40 epochs. A batch size of 128 is used by default unless stated otherwise.

Experiments settings. All experiments are conducted using distributed training (DDP) across $4\times$ NVIDIA GeForce RTX 4090 GPUs, with automatic mixed precision (AMP) enabled to accelerate training.

A. Image-only diffusion representation

Model architecture. For the image modality, we utilize the latent space DiT model [39] as the pre-trained backbone to extract image diffusion representations. We retrieve the DiT-XL/2 checkpoint, pretrained on 256^2 ImageNet from its official codebase for class-conditional generation. We employ it in an unconditional manner by setting the label to null [40]. DiT-XL/2 has 28 Transformer layers, a hidden size of 1152, and 16 attention heads, following the largest configuration of the DiT model family. The off-the-shelf VAE [41] model for latent compression has a down-sample factor of 8, retrieved from Stable Diffusion [42].

To systematically analyze the behavior of intermediate activations in image diffusion models, we perform ablation studies by extracting features from different diffusion timesteps and Transformer blocks, and evaluating their performance on multi-label classification under the linear probing setting.

As shown in Fig. 2, classification performance measured by mAP and OP exhibits a consistent unimodal trend along the diffusion timestep dimension, where intermediate timesteps yield superior results compared to early or heavily corrupted states.

Along the depth dimension, we observe a highly consistent pattern across timesteps: features extracted from the 12th Transformer block consistently achieve the best performance compared to other layers. This empirical result suggests that intermediate layers of the diffusion-Transformer strike a favorable balance between semantic abstraction and spatial specificity, making them particularly well suited for multi-label classification.

B. Text-only diffusion representation

Model architecture. For the text modality, we employ the Plaid 1B model [31] as a pre-trained language diffusion

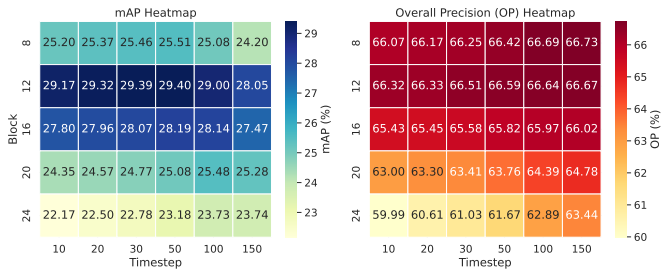


Fig. 2: Multi-label classification performance of image diffusion representations across different Transformer blocks and diffusion timesteps on the VG500. Brighter regions indicate higher mAP (left) or OP (right) values, highlighting the optimal selection of features.

backbone to extract diffusion-based text representations. Plaid 1B is a Transformer-based model with approximately 1.3 billion parameters, whose denoising network consists of 24 Transformer blocks with a hidden dimension of 2048.

For consistency across experiments, the input sequence length is fixed to 600 tokens on VG500 and 150 tokens on Pix2Cap-COCO.

Following the same evaluation protocol as in the image branch, we perform multi-label classification using the extracted text diffusion representations. As illustrated in the fig. 3, representations extracted from deeper Transformer blocks (e.g., block 24) consistently achieve better performance across all diffusion timesteps. Moreover, the best performance is observed at the noise-free timestep ($t = 0$), indicating that early diffusion steps preserve richer semantic information for text-based multi-label classification.

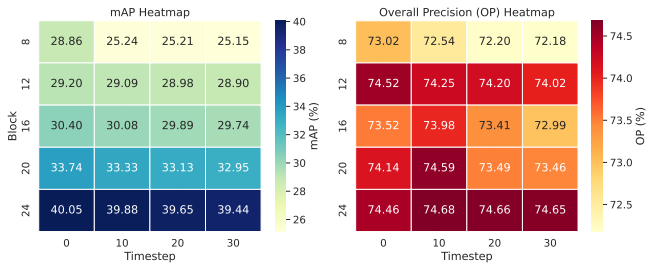


Fig. 3: Multi-label classification performance of text diffusion representations across different Transformer blocks and diffusion timesteps on the VG500. Brighter regions indicate higher mAP (left) or OP (right) values.

C. Cross-modal fusion representation

As discussed in Section III-B, we investigate multiple fusion strategies to combine image and text diffusion representations for multi-label classification.

Specifically, we evaluate four feature fusion methods: *Simple Concat*, *Linear Concat*, *Linear Addition*, and *Cross Attention*. In all methods, image and text features are first ℓ_2-normalized to ensure comparable feature scales across modalities. For *Simple Concat*, the normalized features are directly concatenated. For *Linear Concat*, *Linear Addition*, and *Cross Attention*, the normalized features are further projected into a shared embedding space via modality-specific linear layers before fusion. Unless otherwise specified, the dimensionality of the shared embedding space is set to $d_{\text{alg}} = d_k = 2048$.

Following the ablation study on fusion methods (Table I), we evaluate fusion performance using the *Linear Addition* method with the heuristic block-timestep selection strategy introduced in Section III-B. As shown in Fig. 4, the best performance on VG500 is obtained by fusing image representations extracted at $t_{\text{img}} = 30, b_{\text{img}} = 12$ with text representations extracted at $t_{\text{txt}} = 0, b_{\text{txt}} = 24$, achieving a mAP of 45.71%.

TABLE I: Ablation study of fusion methods on Pix2Cap-COCO ($t_{\text{img}} = 50, b_{\text{img}} = 12; t_{\text{txt}} = 0, b_{\text{txt}} = 24$).

Fusion Method	mAP	CP	CR	CF1	OP	OR	OF1
Simple Concat	78.0	**85.5**	59.9	69.5	**91.6**	67.8	77.91
Linear Concat	79.0	82.8	68.0	74.2	89.0	74.9	81.3
Cross Attention	76.8	83.6	64.9	72.5	89.9	72.1	80.0
Linear Add	**79.6**	82.9	**68.2**	**74.4**	89.2	**75.1**	**81.6**

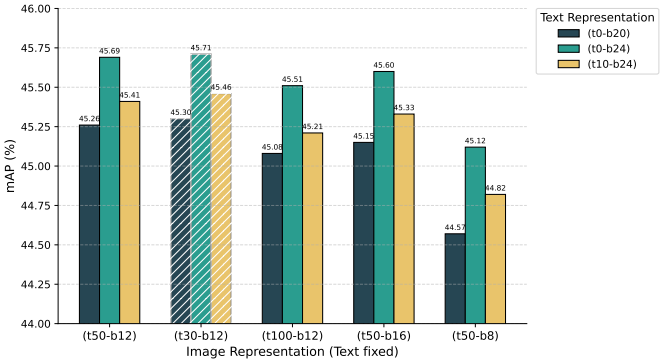


Fig. 4: Multi-label classification performance (mAP) on VG500 using *Linear Addition* fusion across different image–text representation pairs. Each group denotes an image representation (e.g., `(t50-b12)`), and each bar corresponds to a text representation. The best combination in each group is highlighted with white diagonal stripes.

We report the multi-label classification metrics for the optimal block–timestep combinations using Linear Addition fusion strategy, benchmarked against strong baselines on the VG500 (Table II) and Pix2Cap-COCO (Table III), where **"LP"** denotes linear probing, i.e., training a linear classifier on frozen features, and **"FT"** denotes fine-tuning the image diffusion model for feature extraction, after which a classifier is trained on the resulting features.

For VG500, the classifier is first trained for 20 epochs with a learning rate of 1×10^{-3}, followed by joint fine-tuning of the classifier and diffusion model for 2 epochs with a learning rate of 1×10^{-4}. For Pix2Cap-COCO, the initial classifier training

also uses 20 epochs with a learning rate of 1×10^{-3}, while joint fine-tuning is conducted for 5 epochs with a learning rate of 5×10^{-5}. In all fine-tuning experiments, the batch size is set to 16.

The optimal LP configuration on VG500 is $(t_{\text{img}} = 30, b_{\text{img}} = 12, t_{\text{txt}} = 0, b_{\text{txt}} = 24)$, while FT uses $(t_{\text{img}} = 50, b_{\text{img}} = 16, t_{\text{txt}} = 0, b_{\text{txt}} = 24)$. On Pix2Cap-COCO, the best LP setting is $(t_{\text{img}} = 50, b_{\text{img}} = 12, t_{\text{txt}} = 0, b_{\text{txt}} = 24)$, while FT uses $(t_{\text{img}} = 50, b_{\text{img}} = 16, t_{\text{txt}} = 0, b_{\text{txt}} = 24)$.

TABLE II: Comparison with strong baseline models on VG500.

Method	mAP(%)
ResNet-101 [43]	30.9
ResNet-SRN [44]	33.5
SS-GRL [45]	36.6
C-Tran [46]	38.4
DRGN [47]	39.8
DATran [48]	40.1
SADCL [49]	40.5
Q2L-TResL-22k [3]	42.5
ResNet-50 [43] + LLaMA 3.2 1B [50] (LP)	40.5
CLIP (LP) [51]	42.9
ViLT (LP) [52]	37.3
Diff-Feat (Ours, LP)	45.7
Diff-Feat (Ours, FT)	**46.1**

TABLE III: Comparison with strong baseline models on Pix2Cap-COCO.

Method	mAP	CP	CR	CF1	OP	OR	OF1
ResNet-50 + LLaMA3.2 1B (LP)	70.4	73.9	58.1	64.6	82.3	65.9	73.2
CLIP (LP)	76.3	**84.4**	57.8	67.1	**91.8**	65.9	76.7
ViLT (LP)	73.0	84.0	56.9	66.6	90.7	64.0	75.1
Diff-Feat (Ours, LP)	79.6	82.9	**68.2**	74.4	89.2	**75.1**	**81.6**
Diff-Feat (Ours, FT)	**80.2**	83.7	68.0	**74.7**	89.5	75.0	**81.6**

D. Empirical observation: modality-specific trends in diffusion representations

Notation. Let $\mathbf{x}$ denote an input instance(either an image or text embedding). We apply a pre-trained diffusion model to $\mathbf{x}$ via the forward process, yielding latent states $\mathbf{z}_t$ at timestep t (through Eq. 1). Let $\mathbf{h}_{t,b}$ denote the hidden representation extracted from the b-th Transformer block when $\mathbf{z}_t$ is input. Define $\mathcal{A}(t, b)$ as downstream tasks performance (e.g., mAP) using $\mathbf{h}_{t,b}$ as features for linear probing. Empirical findings reveal modality-specific trends in $\mathcal{A}(t, b)$ (Fig. 2 and 3):

(1) Image modality, fixed b: $\mathcal{A}(t, b)$ is unimodal in t; there exists $t^*(b)$ such that

$$\mathcal{A}(t, b) \text{ increases for } t < t^*(b), \quad \text{decreases for } t > t^*(b).$$

(2) Image modality, fixed t: $\mathcal{A}(t, b)$ is unimodal in b; there exists $b^*(t)$ such that

$$\mathcal{A}(t, b) \text{ increases for } b < b^*(t), \quad \text{decreases for } b > b^*(t).$$

(3) Text modality, fixed b: $\mathcal{A}(t, b)$ decreases monotonically with $t \in \mathcal{T}$.

(4) Text modality, fixed t: $\mathcal{A}(t, b)$ increases monotonically with $b \in \mathcal{B}$.

As shown in Fig. 5, for image data, the optimal diffusion timestep tends to vary across a broad range, with local fluctuations. In contrast, the optimal Transformer block is consistently fixed at **layer 12**. Although this may partially relate to the DiT model architecture (which contains 28 layers), it appears remarkably invariant across downstream tasks, dataset distributions, and evaluation metrics. This consistent pattern may offer insights into the internal workings of black-box diffusion models, and motivates the subsequent experimental and mechanism-level analyses presented in Sections IV-E and IV-F.

E. Scaling Analysis of Optimal Layer Position in Vision Diffusion Models

We further evaluate diffusion-Transformer models with different architectural configurations (U-ViT [53] variants) to assess the robustness of the "**Middle Layer**" phenomenon. As shown in Table IV, similar trends persist across model depths, with optimal performance consistently occurring in intermediate layers, demonstrating that this behavior generalizes across architectures. Moreover, we observe a *strong positive correlation* between the optimal layer index and the total number of layers (Spearman $\rho = 0.97$, $p < 0.0015$), indicating that deeper models tend to shift the optimal layer proportionally deeper.

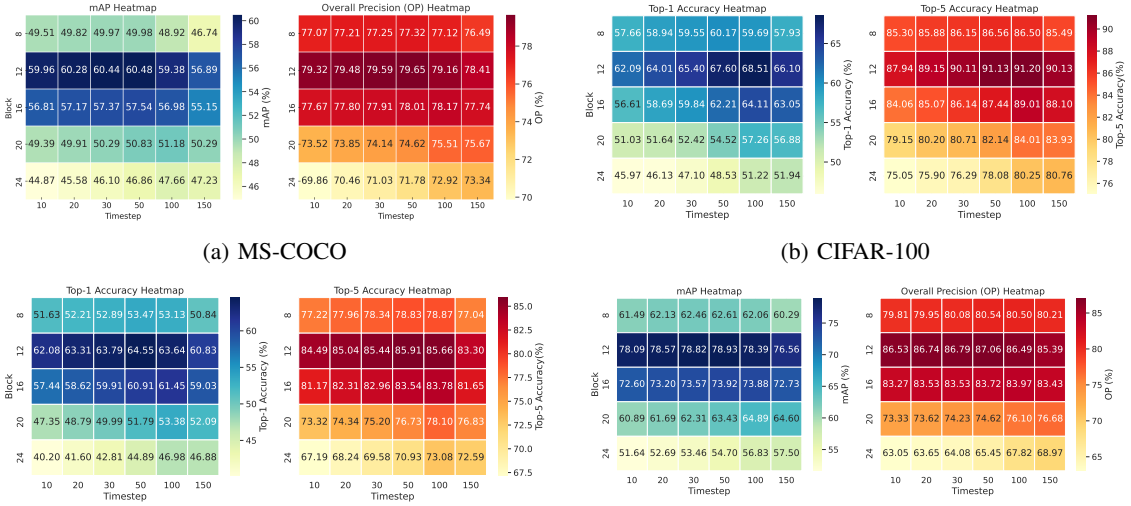
In the U-ViT family, "S", "L", and "H" denote small, large, and huge variants, respectively, while "/2" and "/4" indicate different input patch size. The "Deep" variant refers to models with increased Transformer depth.

TABLE IV: Optimal Layer Selection Across Different Architectures at Diffusion Timestep $t = 50$ on the MS-COCO dataset.

Model Architecture	Optimal Layer	Total Layers	mAP
CIFAR10 (U-ViT-S/2)	10	13	28.24
CelebA 64x64 (U-ViT-S/4)	9	13	27.64
ImageNet 256x256 (U-ViT-L/2)	12	21	59.15
ImageNet 512x512 (U-ViT-L/4)	11	21	54.88
ImageNet 256x256 (U-ViT-H/2)	16	29	64.32
ImageNet 512x512 (U-ViT-H/4)	16	29	61.16

F. Understanding Modality-Dependent Optimality in Diffusion Models

Vision: semantic granularity versus noise robustance. To investigate how semantic granularity manifests in the diffusion representation, we analyze robustness along two orthogonal axes: the diffusion timestep t and the Transformer block depth b. For the timestep axis, we normalize each task's performance by its peak value $r_{\text{task}}(t) = \frac{\text{mAP}_{\text{task}}(t)}{\max_t \text{mAP}_{\text{task}}(t)}$, and define the *$\alpha$-retention interval* as $\mathcal{T}_\alpha(\text{task}) = \{t \mid r_{\text{task}}(t) \geq \alpha\}$, where $\alpha = 0.95$ denotes the range of diffusion timesteps that preserve at least 95% of the task's peak mAP. Along the depth axis, an analogous normalization is applied over Transformer blocks, $r_{\text{task}}(b) = \text{mAP}_{\text{task}}(b) / \max_b \text{mAP}_{\text{task}}(b)$, allowing us



(a) MS-COCO

(b) CIFAR-100

(c) Tiny-ImageNet

(d) PASCAL VOC 2007

Fig. 5: Heatmap visualizations of classification performance across diffusion timesteps and Transformer blocks on multiple image benchmarks.

to compare how coarse- and fine-grained semantics emerge across layers.

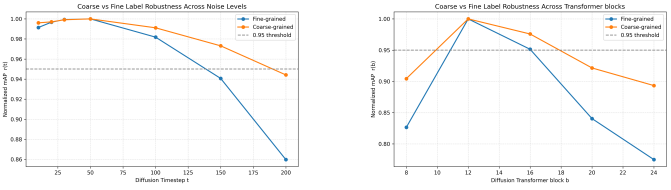
Experiments are conducted on MS-COCO using object labels (fine-grained) and super-category labels (coarse-grained), with linear probes trained on representations extracted from block $b = 12$ of DiT-XL/2-256×256.

As shown in Fig. 6, the fine-grained task drops below the 0.95 threshold at $t = 150$ ($r_{\text{fine}}(150) = 0.941$), yielding $[10, 150] \geq \mathcal{T}_{0.95}(\text{fine}) \geq [10, 100]$. In contrast, the coarse-grained task remains above the threshold even under strong noise corruption ($r_{\text{coarse}}(150) = 0.973$), yielding $[10, 300] \geq \mathcal{T}_{0.95}(\text{coarse}) \geq [10, 200]$. Thus, $|\mathcal{T}_{0.95}(\text{coarse})| \gg |\mathcal{T}_{0.95}(\text{fine})|$.

A complementary pattern emerges along the *depth* axis: when fixing t and sweeping over blocks b, fine-grained labels exhibit a sharp performance peak in the mid-to-late layers (e.g., around block 12), whereas coarse-grained labels remain comparatively stable across a broader span of blocks. Together, the t-axis and b-axis analyses reveal a consistent semantic hierarchy: fine-grained labels rely on detail-sensitive, higher-frequency representations that emerge in later blocks and are easily disrupted by noise, while coarse-grained labels rely primarily on lower-frequency structural cues that persist under strong corruption and across many layers.

Language: semantic neighborhood degradation under noise perturbation. To quantify how diffusion noise affects the local semantic structure of text embeddings, we analyze neighborhood relationships across diffusion timesteps at a fixed Transformer block depth. Let $h(x_t^i) \in \mathbb{R}^{d_{\text{txt}}}$ denote the embedding of sample x^i at diffusion timestep t, and let $N_t(i) = \text{kNN}(h(x_t^i))$ denote the set of k nearest neighbors under cosine similarity, where $k = 10$.

We consider two complementary neighborhood-based metrics to characterize geometric and semantic stability. *Neighborhood overlap* measures geometric stability and is defined as $\text{Overlap}(t) = \sum_{i=1}^{n} \left[|N_0(i) \cap N_t(i)|/k\right]$. *Label consistency* measures semantic alignment and is defined as $\text{Consistency}(t) = \sum_{i=1}^{n} \left[\frac{1}{k} \sum_{j \in N_t(i)} \mathbf{1}(y_i \cap y_j \neq \emptyset)\right]$, where y_i denotes the multi-label annotation associated with sample i.



(a) Fixed block $b = 12$.

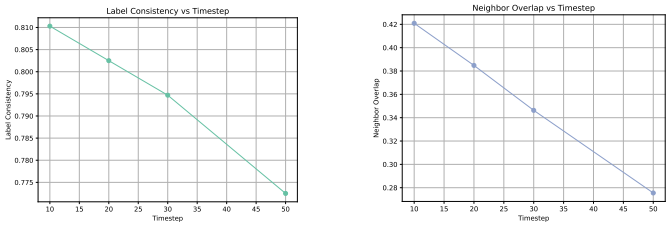
(b) Fixed diffusion timestep $t = 50$.

Fig. 6: Robustness comparison between coarse- and fine-grained labels along two orthogonal axes of DiT on the MS-COCO dataset.

Fig. 7 presents both metrics as functions of the diffusion timestep t at block depth $b = 24$ of Plaid-1B. As t increases, neighborhood overlap decreases monotonically, indicating a progressive disruption of the local embedding geometry. Label consistency also declines smoothly, reflecting a gradual loss of semantic coherence among nearest neighbors. Together, these results demonstrate that diffusion noise systematically degrades both geometric and semantic neighborhood structure of text embeddings, suggesting that, in the textual domain, noise acts as a destructive perturbation rather than a beneficial regularizer.

V. Discussion

For images, adaptive noise levels and Transformer blocks help remove redundant details while preserving task-relevant



(a) Label consistency. (b) Neighborhood overlap.

Fig. 7: Semantic stability of text embeddings under different diffusion timestep at a fixed block depth $b = 24$ on the VG500.

features, leading to a peak in discriminative quality at the intermediate noise level and block [54]; In contrast, text representations are more sensitive to corruption: once corrupted by noise, semantic information becomes difficult to recover [31], resulting in a monotonic degradation with increasing noise level. Meanwhile, deeper transformer blocks use self-attention to better understand text context [55], explaining the upward trend in block depth.

To further analyze the evolution of image representations across diffusion timesteps and Transformer layers, we visualize attention maps from the image diffusion model. Fig. 8 shows that intermediate timesteps and mid-level blocks yield more coherent and semantically focused attention, providing qualitative support for our quantitative findings.

Overall, our findings suggest that intermediate diffusion representations-properly selected in terms of block depth and diffusion timestep-can serve as strong and general-purpose features for downstream multi-label classification tasks.

We attribute the strong performance of the fused representation in highly imbalanced multi-label tasks to the powerful generative capacity of pre-trained diffusion models. In the meantime, the choice of optimal block-timestep pairs and effective fusion strategies plays a crucial role.

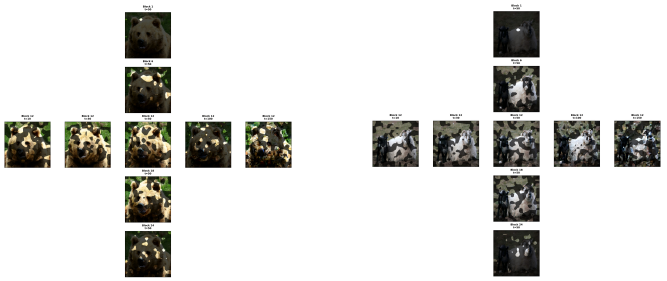


Fig. 8: Attention map visualization across diffusion timesteps and Transformer blocks for two representative samples in the MS-COCO. The maps demonstrate how the model gradually focuses on semantically meaningful regions.

VI. Conclusions and Future Work

In this paper, we propose *Diff-Feat*, a simple yet effective framework for multi-label classification. By identifying optimal block-timestep combinations from image and text diffusion representations and leveraging a heuristic search strategy with a lightweight fusion mechanism, our method achieves strong performance, reaching 46.1% mAP on VG500 and 80.2% mAP on Pix2Cap-COCO, outperforming competitive baseline models. In addition, we provide new insights and mechanism-level analyses into the varying effectiveness of different block-timestep configurations across downstream tasks. Meanwhile, the scaling analysis further deepens our understanding of the representation properties of diffusion-Transformer models across different architectures, revealing the stability and consistency of optimal intermediate representations as model depth varies. We believe that *Diff-Feat* offers a generalizable and adaptable solution for a wide range of multi-label classification scenarios, with potential applications as a strong self-supervised foundation in medical diagnosis and other specialized domains.

Despite its strong empirical performance, *Diff-Feat* has several limitations. First, although the heuristic search strategy significantly reduces computational cost, it may overlook globally optimal fusion configurations, particularly in more complex settings. Second, as research on diffusion-based language models continues to advance, discrete diffusion language models have demonstrated strong generative performance [12]. However, the present work is limited to continuous diffusion language models, which constitutes an important direction for future research. Moreover, the proposed framework has the potential to be extended to real-world applications such as medical diagnosis, where it may contribute to positive societal impact and provide practical value in high-stakes decision-making scenarios.

Acknowledgments

This work was supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001).

References

[1] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *CoRR*, abs/1807.07247, 2018.

[2] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4657–4666, 2015.

[3] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *CoRR*, abs/2107.10834, 2021.

[4] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2256–2265. JMLR.org, 2015.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural*

Information Processing Systems, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[9] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

[10] Rongjie Huang, Jiawei Huang, Dongchao Yang, et al. Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[11] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[12] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025.

[13] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022.

[14] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15802–15812, October 2023.

[15] 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 (CVPR)*, pages 2955–2966, June 2023.

[16] Shuchen Xue, Tianyu Xie, Tianyang Hu, Zijin Feng, Jiacheng Sun, Kenji Kawaguchi, Zhenguo Li, and Zhi-Ming Ma. Any-order gpt as masked diffusion model: Decoupling formulation and architecture, 2025.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, May 2017.

[18] Zuyao You, Junke Wang, Lingyu Kong, Bo He, and Zuxuan Wu. Pix2cap-coco: Advancing visual comprehension via pixel-level captioning, 2025.

[19] Ruijie Yao, Sheng Jin, Lumin Xu, Wang Zeng, Wentao Liu, Chen Qian, Ping Luo, and Ji Wu. Gkgnet: Group k-nearest neighbor based graph convolutional network for multi-label image recognition. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 91–107, Cham, 2025. Springer Nature Switzerland.

[20] Shichao Xu, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Zhu Qi. Open vocabulary multi-label classification with dual-modal decoder on aligned visual-textual features, 2023.

[21] Shuping Yuan, Yang Chen, Chengqiong Ye, et al. Cross-modal multi-label image classification modeling and recognition based on nonlinear. *Nonlinear Engineering*, 12(1):20220194, 2023.

[22] Yangtao Wang, Yanzhao Xie, Jiangfeng Zeng, et al. Cross-modal fusion for multi-label image classification with attention mechanism. *Computers and Electrical Engineering*, 101:108002, 2022.

[23] Shuyi Ouyang, Hongyi Wang, Ziwei Niu, et al. Hsvlt: Hierarchical scale-aware vision-language transformer for multi-label image classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4768–4777, New York, NY, USA, 2023. Association for Computing Machinery.

[24] David Sebastian Hoffmann, Kai Norman Clasen, and Begüm Demir. Transformer-based multi-modal learning for multi-label remote sensing image classification. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4891–4894, 2023.

[25] Runhui Huang, Jianhua Han, Guansong Lu, et al. DiffDis: Empowering Generative Diffusion Model with Cross-Modal Discrimination Capability . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15667–15677, Los Alamitos, CA, USA, October 2023. IEEE Computer Society.

[26] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 899–907, Red Hook, NY, USA, 2013. Curran Associates Inc.

[27] 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*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

[28] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, et al. Do text-free diffusion models learn discriminative visual representations? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LX*, page 253–272, Berlin, Heidelberg, 2024. Springer-Verlag.

[29] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F. Polanía, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: stable diffusion complements dino for zero-shot semantic correspondence. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[30] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Seo Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[31] Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[32] Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27558–27568, June 2024.

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[34] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 522–531, 2019.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[36] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[37] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015. CS231n Course Project Report, Stanford University.

[38] Mark Everingham, S. M. Eslami, Luc Gool, et al. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, January 2015.

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023.

[40] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[44] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2027–2036, 2017.

[45] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 522–531, 2019.

[46] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16473–16483, 2021.

[47] Wei Zhou, Weitao Jiang, Dihu Chen, Haifeng Hu, and Tao Su. Mining semantic information with dual relation graph network for multi-label image classification. *IEEE Transactions on Multimedia*, 26:1143–1157, 2024.

[48] Wei Zhou, Zhijie Zheng, Tao Su, and Haifeng Hu. Datran: Dual attention transformer for multi-label image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):342–356, 2024.

[49] Lei Ma, Dengdi Sun, Lei Wang, Hai Zhao, and Bin Luo. Semantic-aware dual contrastive learning for multi-label image classification. *ArXiv*, abs/2307.09715, 2023.

[50] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. The llama 3 herd of models, 2024.

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[52] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

[53] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

[54] Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. *Revelio*: Interpreting and leveraging semantic information in diffusion models, 2024.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.