EVE: Efficient Vision-Language Pre-training with Masked Prediction and Modality-Aware MoE

Junyi Chen1*, Longteng Guo2, Jia Sun3, Shuai Shao3, Zehuan Yuan3, Liang Lin1, Dongyu Zhang1†

1Sun Yat-sen University
2Institute of Automation, Chinese Academy of Sciences (CASIA)
3Bytedance Inc

chenjy765@mail2.sysu.edu.cn, longteng.guo@nlpr.ia.ac.cn, {sunjia.ly, shaoshuai.0516, yuanzehuan}@bytedance.com
linliang@ieee.org, zhangdy27@mail.sysu.edu.cn

Abstract

Building scalable vision-language models to learn from diverse, multimodal data remains an open challenge. In this paper, we introduce an Efficient Vision-language: foundation model, namely EVE, which is one unified multimodal Transformer pre-trained solely by one unified pre-training task. Specifically, EVE encodes both vision and language within a shared Transformer network integrated with modality-aware sparse Mixture-of-Experts (MoE) modules, which capture modality-specific information by selectively switching to different experts. To unify pre-training tasks of vision and language, EVE performs masked signal modeling on image-text pairs to reconstruct masked signals, i.e., image pixels and text tokens, given visible signals. This simple yet effective pre-training objective accelerates training by 3.5x compared to the model pre-trained with Image-Text Contrastive and Image-Text Matching losses. Owing to the combination of the unified architecture and pre-training task, EVE is easy to scale up, enabling better downstream performance with fewer resources and faster training speed. Despite its simplicity, EVE achieves state-of-the-art performance on various vision-language downstream tasks, including visual question answering, visual reasoning, and image-text retrieval.

Introduction

Vision-Language Pre-training aims to learn a general multimodal representation that can be transferred to various vision-language downstream tasks, such as vision-language understanding and image-text retrieval. A vision-language foundation model should have excellent performance while being easy to train and scale up, which can be achieved through the model architecture and the pre-training tasks.

The model architectures of recent methods can be roughly divided into two categories: dual-encoder architecture and unified architecture. Dual-encoder methods (Radford et al. 2021; Zeng, Zhang, and Li 2022) employ modality-specific models (e.g., BERT (Devlin et al. 2019), ViT (Dosovitskiy et al. 2021)) to encode different modalities separately and a fusion module to integrate them. As for the fusion module, some methods (Radford et al. 2021) employ shallow fusion (e.g., dot product) for the interaction of vision and language. Some alternative methods (Zeng, Zhang, and Li 2022) use deep neural networks, such as Transformer Encoders, to perform deep fusion on modality interaction, but lead to difficulties in scaling up and low efficiency. Unified methods (Kim, Son, and Kim 2021; Wang et al. 2022b) use a modality-shared Transformer to encode different modalities jointly. This approach simplifies the framework and improves the speed, helping with model scaling up. However, they overlook the inherent gap between modalities, leading to lower overall performance. Image is continuous, redundant, and low-level on the raw signals, while text is discrete, refined, and high-level. Directly using a shared Transformer to encode different modalities with semantic gap poses problems. Therefore, it is necessary to consider the differences between different modalities carefully.

Previous methods also have explored numerous pre-training tasks for vision-language pre-training, including Image-Text Contrastive Learning (Radford et al. 2021), Image-Text Matching (Li et al. 2021), Word-Patch Alignment (Chen et al. 2020), Masked Language Modeling (Su et al. 2020), Masked Image Modeling (Bao et al. 2022b), and so on. They have been widely used to improve vision-language pre-training. While incorporating more pre-
training tasks can enhance performance, adding too many tasks can also lead to some problems. Foremost, it significantly prolongs the pre-training time and increases the computational resources required. Additionally, it necessitates manual weight adjustments for different objectives. Furthermore, excessive pre-training objectives can result in a reduction in the model’s scalability, which is crucial in designing pre-training models, as the recent success has shown in large language models (Ouyang et al. 2022; Wei et al. 2022b; Zhao et al. 2023b). Therefore, it is necessary to use effective and scalable pre-training tasks.

In this paper, we propose an Efficient Vision-language foundation model (EVE) with a unified modality-aware Transformer pre-trained with a single unified pre-training task, i.e., masked signal modeling.

In terms of model architecture, we use a unified modality-aware Transformer and revisit the integration of Mixture-of-Experts in vision-language pre-training. We employ a shared Multi-Head Self-Attention module and a Modality-Aware MoE module for the modality-aware Transformer to encode and fuse various modalities jointly. Using a unified shared Transformer is more concise and flexible, which simplifies the extension to additional modalities and facilitates cross-modal alignment. By incorporating MoE, we can take into account the differences between modalities and capture more modality-specific information. We also introduce a modality routing technique in MoE that enables the router select more appropriate experts for processing.

In terms of pre-training tasks, we propose a unified masked signal modeling technique combining masked pixel and language modeling, which significantly improves training speed and reduces scaling difficulty. Some methods (Wang et al. 2023; Kwon et al. 2023; Zhao et al. 2023a) have applied generative pre-training paradigm to vision-language pre-training. While they either add the generative objective with other complex objectives like ITC and ITM (Kwon et al. 2023) or employ more complicated targets such as visual tokens (Wang et al. 2023) or momentum features (Zhao et al. 2023b), which require a nontrivial visual tokenizer or momentum model. All of these increase the complexity of pre-training. In contrast to them, we just utilize the raw signals from the image-text pairs themselves to minimize the complexity of pre-training and achieve better scalability. Pre-training speed is 3.5x faster than incorporating ITC and ITM.

EVE can greatly enhance pre-training speed, as shown in Figure 1. It decreases the demand for extensive computational resources while being easy to scale up. We demonstrate the effectiveness of EVE on various vision-language downstream tasks, including visual question answering, visual reasoning, and image-text retrieval. EVE achieves state-of-the-art performance on Image-Text Retrieval and Vision-Language Understanding (VQA and NLVR2) tasks.

Our contributions are summarized as follows:

- We introduce EVE, an efficient vision-language foundation model that achieves state-of-the-art performance while improving training speed, with one unified multimodal Transformer and one unified pre-training task.
- We integrate Modality-Aware MoE with a shared multi-modal Transformer to achieve a more profound fusion of different modalities and capture more modality-specific information simultaneously, resulting in better performance and faster inference speed within a unified architecture.
- We propose a unified masked signal modeling technique, simplifying vision-language pre-training into a single unified objective, resulting in significantly improved pre-training speed and competitive performance.

### Related Work

Model architecture and pre-training tasks are crucial factors in the representation learning of vision-language.

#### Model Architecture

- **Dual-encoder with a fusion module** (Li et al. 2021; Liu et al. 2021; Dou et al. 2022b; Zhao et al. 2023a) performs well on vision-language tasks but with higher time and architecture complexity. Unified architecture methods (Kim, Son, and Kim 2021; Wang et al. 2022b; Bao et al. 2022a,b) can flexibly encode different modalities as a fusion encoder or process a single modality as a unimodal encoder, demonstrating faster inference speed and promising performance. Some of them (Kim, Son, and Kim 2021; Wang et al. 2022b) use a shared standard Transformer (Vaswani et al. 2017) to jointly encode different modalities, while they ignore the modality gap and lead to worse performance. Others (Bao et al. 2022a,b) use MoME Transformer instead and prove that shared attention is better for multimodal learning. However, MoME Transformer uses modality-shared FFN in the deep layers may neglect some modality-specific information.

Considering the simplicity, effectiveness, and flexibility of the unified architecture, we adopt a unified architecture with Modality-Aware MoE to better capture modality specifics during fusion for multimodal representation learning. We achieve state-of-the-art performance with approximately the same inference cost.

#### Masked Signal Modeling

Recently, several methods (Bao et al. 2022b; Zhao et al. 2023a; He et al. 2022b; Geng et al. 2022) explore the "mask then predict" paradigm in the vision for vision-language pre-training. While VLBEiT (Bao et al. 2022b) introduces training on the visual modality through masked image modeling, their reconstruction target is the visual token, which may significantly influence performance depending on the visual tokenizer employed. DAVINCI (Diao et al. 2023) extends prefix language modeling further to vision, but it also uses the discrete visual token as the target. MAMO (Zhao et al. 2023a) enriches multimodal representation by using momentum features in masked representation modeling, which relies heavily on a momentum teacher model to avoid divergence. Some methods (Kwon et al. 2023; He et al. 2022b; Gui et al. 2023) use masked pixel modeling, but they all require additional costly pre-training tasks such as ITC and ITM (Li et al. 2019). Among these methods, VL-MAE (He et al. 2022b) only applies masked pixel modeling to the image encoder. M3AE (Geng et al. 2022) leverages a
unified Image-Language masking approach to mask and reconstruct both images and text simultaneously, but it is not used in multimodal downstream tasks.

We unify masked pixel and language modeling into masked signal modeling, reconstructing masked raw signals from visible signals. This simplifies and accelerates training, achieving better performance and scalability.

**Mixture-of-Experts (MoE)** Mixture-of-Experts has been extensively explored in computer vision (Shazeer et al. 2017; Lepikhin et al. 2021). These methods generally aim to improve performance by learning a better routing using auxiliary losses (Lepikhin et al. 2021; Zoph et al. 2022), converting it into a linear assignment problem (Lewis et al. 2021), or making it differentiable (Hazimeh et al. 2021). MoE seems well-suited for multimodal learning, but the differences between modalities present some challenges. LMoE (Mustafa et al. 2022) involves more auxiliary losses to balance different modalities, uni-perceiver-moe (Zhu et al. 2022) employs conditional MoE, VLMO (Bao et al. 2022a) and VLMoE (Shen et al. 2023) use shared expert in the deep layers.

However, existing methods increase complexity or limit performance due to manual routing and ignoring modality information. Therefore, we propose Modality-Aware MoE as a simple way to apply MoE to multimodal learning. We simplify the auxiliary loss and capture more modality specifics by expert switching.

**Methods**

**Backbone Network**

As shown in Figure 2, we adopt a unified multimodal Transformer with shared attention and Modality-Aware Mixture-of-Experts as the backbone network, which is capable of encoding different modalities. After pre-training, the model can be utilized as either a fusion encoder or a unimodal encoder for various downstream tasks through fine-tuning.

For Image $I$, following ViT (Dosovitskiy et al. 2021), we first split the Image $I$ into $N$ patches with a patch size of $P$. The resulting $N = HW/P^2$ patches are projected into a shared embedding space using a linear projector. A special token $I_{cls}$ is added at the beginning of all visual tokens. We employ learnable visual position embeddings $I_{pos}$ and visual type embeddings $I_{type}$ on visual tokens. Image embedding can be summarized as follows.

$$I_{emb} = [I_{cls}, I_1, \ldots, I_N] + I_{pos} + I_{type} \quad (1)$$

For Text $T$, following BERT (Devlin et al. 2019), we tokenize the text into discrete tokens with the maximum length of $n$ and project them into the joint embedding space. We add a special token $T_{cls}$ at the beginning of all text tokens and use learnable text position embeddings $T_{pos}$ and text type embeddings $T_{type}$ for text encoding. Text embedding can be summarized as follows.

$$T_{emb} = [T_{cls}, T_1, \ldots, T_n] + T_{pos} + T_{type} \quad (2)$$

We concatenate $I_{emb}$ and $T_{emb}$ as the input to the model:

$$P_{emb} = [I_{emb}, T_{emb}] \quad (3)$$

**Modality-Aware Mixture-of-Experts**

Multimodal learning differs significantly from unimodal learning, as the differences between modalities cannot be ignored. Using the same Feed-Forward Network for all modalities can lead to inappropriate fusion of modalities, resulting in degraded performance. Conversely, using modality-specific MoE in all layers may not benefit the alignment of different modalities. Therefore, we propose the Modality-Aware Mixture-of-Experts (MoE) as shown in Figure 3, which incorporates the modality routing technique on top
of the general MoE to capture modality-specific information while fusing by selectively switching to different experts. In the general MoE, each MoE block typically consists of \( N \) experts, and each input token is processed by \( k \) experts selected from the \( N \) experts. A lightweight router \( g \) is used to select the \( k \) experts for each token, which employs a simple linear-softmax predictor to calculate the routing weight. This can be formulated as:

\[
g(x) = \text{softmax}(W \cdot x) \quad (4)
\]

\( W \in \mathbb{R}^{D \times N} \) is a learnable projector for input \( x \in \mathbb{R}^D \).

The final output of the MoE block is the weighted average of the \( k \) selected experts, which can be formulated as:

\[
\text{MoE}(x) = \sum_{i=1}^{k} g(x)_i \cdot \text{FFN}_i(i) \quad (5)
\]

**Modality Routing**

General MoE does not impose any restrictions on the router, which can easily lead to unbalanced routing. LIMoE (Mustafa et al. 2022) points out that this phenomenon can be exacerbated in multimodal learning due to the difference in token count across different modalities.

To address this issue, we propose a modality-aware routing approach to enhance the router. We adopt a best-effort strategy for routing to preserve all tokens while explicitly providing modality information to the router by adding modality-specific embeddings. The new routing function can be formulated as follows:

\[
g(x) = \text{softmax}(W \cdot (x + b_m)) \quad (6)
\]

Here, we use modality-specific embeddings \( b_m \in \mathbb{R}^D \) for different modalities, i.e., \( b_I \) for images and \( b_T \) for text.

**Auxiliary Loss**

In addition to modality routing, we use a simple single auxiliary loss to balance routing and avoid carefully tuning the weight. Following Shazeer et al. (2017), we add Load-Balancing Loss as the auxiliary loss to train the router. It can be formulated as follows:

\[
\mathcal{L}_{aux} = \alpha \cdot N \sum_i^N f_i \cdot p_i \quad (7)
\]

This objective encourages uniform routing of tokens, where \( N \) denotes the number of experts, \( f_i \) denotes the fraction of tokens dispatched to the \( i \)th expert, and \( p_i \) denotes the average routing weight for the \( i \)th expert. The weight \( \alpha \) is a hyperparameter that we set at 0.001 by default to avoid overwhelming other objectives.

Considering efficiency, we use a soft router with top-\( k = 2 \) in the deep layers and a hard router in the shallow layers. An MoE module equipped with a hard router has the same number of experts as the number of modalities. The hard router directly selects the corresponding expert based on the modality of each token.

**Pre-training Task: Masked Signal Modeling**

Previous multimodal models (Li et al. 2021; Radford et al. 2021; Bao et al. 2022a; Li et al. 2019; Zhao et al. 2023a) typically involve complex pre-training tasks like Image-Text Contrastive Learning (ITC) (Radford et al. 2021), Image-Text Matching (ITM) (Li et al. 2019), and Masked Representation Modeling (MRM) (Zhao et al. 2023a). These methods have shown good performance, but pre-training still requires significant computational resources, and is challenging to scale up.

Table 1 shows the efficiency comparison between different pre-training tasks, which indicates a significant difference in time consumption and batch size. Compared to pre-training without ITC and ITM, including them requires four times more computational resources to achieve a similar speed. Moreover, ITC and ITM tasks are similar to other contrastive learning-based methods that typically require a larger batch size to achieve better performance. Incorporating additional pre-training tasks can significantly decrease training speed, increase training difficulty, and have an impact on the scalability of the model.

Thus, we pre-train our model with only one unified masked signal modeling objective on image-text pairs to reconstruct masked signals by visible signals as shown in Figure 2. Specifically, masked signal modeling combines masked image modeling and masked language modeling, and only utilizes the raw signals from image-text pairs themselves without relying on any additional techniques. We use masked image and complete text in masked image modeling, while complete image and masked text in masked language modeling. Despite its simplicity, our ap-

<table>
<thead>
<tr>
<th>Pre-training Tasks</th>
<th>MLM</th>
<th>ITC</th>
<th>ITM</th>
<th>MIM Token</th>
<th>MIM Pixel</th>
<th>Batch size</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITC</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>224</td>
<td>2.14h</td>
</tr>
<tr>
<td>ITM</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>152</td>
<td>3.09h</td>
</tr>
<tr>
<td>MIM Token</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>132</td>
<td>3.26h</td>
</tr>
<tr>
<td>MIM Pixel</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>80</td>
<td>6.88h</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64</td>
<td>7.73h</td>
</tr>
</tbody>
</table>

Table 1: Maximum batch size per GPU and pre-training time per epoch of different pre-training tasks on 8 A100 GPUs with the same architecture as EVE-Base. We add vision mask tokens in the encoder during masked token modeling.
proach achieves competitive performance compared to previous methods and can be easily scaled up.

In this section, we use $h(\cdot)$ and $\theta(\cdot)$ to denote the encoder and the decoder. $\hat{I}$ and $\hat{T}$ are represented for masked image and masked text. $D$ indicates the dataset.

**Masked Language Modeling (MLM)** Following BERT (Devlin et al. 2019), we randomly mask some of the text tokens and predict them based on the information provided by the image and corrupted text. The Masked Language Modeling (MLM) objective can be formulated as follows:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(I, T) \sim D} l_{\text{mlm}}(\theta_{\text{t}}(h(I, \hat{T})), T)$$

$\ell_{\text{mlm}}$ computes the cross-entropy loss between the prediction probability $P_{\text{mlm}}$, obtained from the text decoder $q_{\text{t}}$, and the ground truth on each masked token. We use a two-layer MLP with a softmax layer as the text decoder.

**Masked Image Modeling (MIM)** Previous methods (Zhao et al. 2023a; Zhang et al. 2023; Wang et al. 2023) have typically employed semantically rich visual features obtained from the model itself or discrete visual tokens obtained from visual tokenizers as the targets for MIM. However, both approaches have their drawbacks. Training visual tokenizers (Ramesh et al. 2021; Peng et al. 2022) is a challenging task as different tokenizers can have varying impacts on performance and may lead to error propagation. Meanwhile, using visual features (Zhao et al. 2023a; Zhang et al. 2023) requires either applying momentum distillation techniques or employing other loss functions and techniques to prevent the model from diverging during training. These MIM targets make the overall framework more complex.

In visual self-supervised learning, some works use other information as the MIM targets, such as RGB pixels (He et al. 2022a), scene depth (Bachmann et al. 2022), HOG (Wei et al. 2022a), etc. However, using targets such as scene depth and HOG requires additional techniques, which increases the complexity of the training process. In order to maintain simplicity and effectiveness, we choose to utilize the image pixels themselves as the reconstruction target.

Following MAE (He et al. 2022a), we adopt an asymmetric design for MIM, where only observed image patches and all text tokens are fed into the encoder. A lightweight decoder is used to reconstruct raw pixels on masked positions from partial image representation and masked tokens, as shown in Figure 2. We use multiple Transformer blocks with narrower hidden widths as the decoder. The MIM objective can be formulated as:

$$\mathcal{L}_{\text{mim}} = \mathbb{E}_{(I, T) \sim D} l_{\text{mim}}(\theta_{\text{t}}(h(I, \hat{T})), I)$$

$l_{\text{mim}}$ calculates the mean square error between the raw pixels and the reconstructed result generated by the image decoder. We compute the loss on masked image patches.

The overall objective of masked signal modeling is:

$$\mathcal{L} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{mim}}$$

### Experiments

**Pre-training Datasets**

Following Previous methods, we pre-train EVE on four widely used public datasets: MSCOCO Captions (Lin et al. 2014), Visual Genome (Krishna et al. 2017), SBU Captions (Ordonez, Kulkarni, and Berg 2011) and Conceptual Captions (Sharma et al. 2018). There are about 4M images and 10M image-text pairs in all datasets. Since some downstream tasks are based on COCO, we exclude all images in the test sets of downstream tasks from the pre-training data. We also pre-train EVE-Large on a larger dataset with 21M image-text pairs by adding CC12M (Changpinyo et al. 2021).

**Implementation Details**

EVE-Base has 12 Transformer blocks and EVE-Large has 24 Transformer blocks. We employ a soft router with 32 experts in EVE-Base on top-2 blocks, EVE-Large on top-3 blocks, and a hard router on the other blocks. We pre-train EVE-Base for 480k steps with a batch size of 2048 and EVE-Large with the same batch size for 280k steps. We use AdamW (Loshchilov and Hutter 2019) optimizer. The peak learning rate is 5e-4 for EVE-Base and 2e-4 for EVE-Large. During pre-training, the image resolution is $224 \times 224$. We use random resized cropping and horizontal flipping for data augmentation. We mask 75% of image in MLM and 50% of text in MIM. EVE is initialized with BEiTv2. More details are provided in Appendix 1.

**Vision-Language Downstream Tasks**

We evaluate our pre-trained model on three common Vision-Language Tasks. More implementation details and comparison on inference speed are provided in Appendix.

**Visual Question Answering (VQA)** VQA requires the model to predict an answer based on the given image and question. We use VQA2.0 dataset (Goyal et al. 2017) to evaluate our model. Following previous work (Bao et al. 2022a), we view the task as a classification task.

**Natural Language for Visual Reasoning (NLVR2)** Given a sentence and two images, NLVR2 asks the model to judge whether the sentence accurately describes the relationship between the two images. We evaluate our model on NLVR2 dataset (Suhr et al. 2019). Following Chen et al. (2020), we convert the triplet input into two image-text pairs with the same text description and different images.

**Image-Text Retrieval** Retrieval task contains two sub-tasks: Image-to-Text Retrieval (TR) and Text-to-Image Retrieval (IR). We evaluate the model on widely used Flickr30K (Plummer et al. 2015) and MSCOCO (Lin et al. 2014) benchmarks following Karpathy split (Karpathy and Fei-Fei 2015). Following Li et al. (2021), we apply ITC and ITM losses in the fine-tuning stage and we use rerank strategy during inference.

---

1Full manuscript with appendix is in arXiv:2308.11971.
Table 2: Comparison with state-of-the-art base-size models on VQA, NLVR2, MSCOCO and Flickr30K. Gray lines indicate the model pre-trained with much more data (more than 400M).

<table>
<thead>
<tr>
<th>Model</th>
<th>#Images</th>
<th>VQA test-dev</th>
<th>VQA test-std</th>
<th>NLVR2 dev</th>
<th>NLVR2 test-P</th>
<th>COCO TR@1</th>
<th>COCO IR@1</th>
<th>Flickr30K TR@1</th>
<th>Flickr30K IR@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVE-Base (Ours)</td>
<td>4M</td>
<td>78.00</td>
<td>78.02</td>
<td>83.34</td>
<td>83.93</td>
<td>79.6</td>
<td>62.0</td>
<td>95.6</td>
<td>86.4</td>
</tr>
<tr>
<td>EVE-Large16M (Ours)</td>
<td>16M</td>
<td>76.33</td>
<td>76.44</td>
<td>80.95</td>
<td>82.10</td>
<td>76.3</td>
<td>62.6</td>
<td>95.6</td>
<td>84.0</td>
</tr>
<tr>
<td>EVE-Large4M (Ours)</td>
<td>4M</td>
<td>75.45</td>
<td>75.40</td>
<td>81.58</td>
<td>81.98</td>
<td>76.3</td>
<td>60.1</td>
<td>95.6</td>
<td>84.5</td>
</tr>
<tr>
<td>EVE-Large8M (Ours)</td>
<td>8M</td>
<td>74.64</td>
<td>74.89</td>
<td>82.77</td>
<td>83.34</td>
<td>74.8</td>
<td>57.2</td>
<td>92.3</td>
<td>79.3</td>
</tr>
<tr>
<td>EVE-Large16M (Ours)</td>
<td>16M</td>
<td>73.33</td>
<td>73.40</td>
<td>81.93</td>
<td>82.66</td>
<td>72.9</td>
<td>55.1</td>
<td>95.8</td>
<td>83.9</td>
</tr>
<tr>
<td>EVE-Large32M (Ours)</td>
<td>32M</td>
<td>72.12</td>
<td>72.24</td>
<td>81.58</td>
<td>81.98</td>
<td>72.1</td>
<td>53.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large64M (Ours)</td>
<td>64M</td>
<td>71.12</td>
<td>71.24</td>
<td>81.58</td>
<td>81.98</td>
<td>71.1</td>
<td>51.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large128M (Ours)</td>
<td>128M</td>
<td>70.12</td>
<td>70.24</td>
<td>81.58</td>
<td>81.98</td>
<td>70.1</td>
<td>49.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large256M (Ours)</td>
<td>256M</td>
<td>69.12</td>
<td>69.24</td>
<td>81.58</td>
<td>81.98</td>
<td>69.1</td>
<td>47.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large512M (Ours)</td>
<td>512M</td>
<td>68.12</td>
<td>68.24</td>
<td>81.58</td>
<td>81.98</td>
<td>68.1</td>
<td>45.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large1024M (Ours)</td>
<td>1024M</td>
<td>67.12</td>
<td>67.24</td>
<td>81.58</td>
<td>81.98</td>
<td>67.1</td>
<td>43.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large2048M (Ours)</td>
<td>2048M</td>
<td>66.12</td>
<td>66.24</td>
<td>81.58</td>
<td>81.98</td>
<td>66.1</td>
<td>41.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
<tr>
<td>EVE-Large4096M (Ours)</td>
<td>4096M</td>
<td>65.12</td>
<td>65.24</td>
<td>81.58</td>
<td>81.98</td>
<td>65.1</td>
<td>39.0</td>
<td>92.3</td>
<td>78.9</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on MIM target. × denotes divergence during fine-tuning.

Results on Downstream Tasks
We present the results of VQA, NLVR2, COCO, and Flickr30K with state-of-the-art base models in Table 2 and large models in Table 4. We report the accuracy for VQA and NLVR2, top-1 recall for TR and IR.

Results on Vision-Language Understanding
EVE-Base outperforms all previous methods on Understanding tasks and even marginally outperforms BEiT3-Base (Wang et al. 2023) pre-trained with 3.1B data on VQA. EVE-Base outperforms VLMO (Bao et al. 2022a), which also employs a unified architecture with more pre-training objectives by 1.77% on VQA test-dev and 0.70% on NLVR2 test-P. EVE-Large4M shows similar performance to SimVLM-Large (Wang et al. 2022b), whereas EVE-Large16M surpasses SimVLM-Huge which is larger and pre-trained on much more data.

Results on Image-Text Retrieval
EVE-Base achieves competitive results on Flickr and outperforms the previous state-of-the-art methods on COCO. Compared to VLMO, EVE-Base achieves improvements of 6.42% on COCO text retrieval R@1 and 8.59% on COCO image retrieval R@1. In addition, EVE-Large demonstrates better performance on both COCO and Flickr30K to other Large or even Huge models with very limited data. Notably, Image-Text Contrastive Learning and Image-Text Matching are not involved in the pre-training of EVE.

Ablation Studies
For all ablation studies, we pre-train the model for 25 epochs with a similar architecture to EVE-Base and report accuracy on NLVR2, VQA dev set, and top-1 recall on Flickr30K. We use the soft router with top-k = 2 by default. We present some more ablation studies in Appendix.

MIM Target
We compare different MIM targets in Table 3, including image token and pixel. We use the tokenizer from BEiT v2 (Peng et al. 2022) and DALL-E (Ramesh et al. 2021). It is observed that reconstructing pixels is better than reconstructing image tokens in all tasks. Using a more complex MIM target does not achieve the expected effect.

Masking Ratio
In Figure 4, we investigate the impact of different masking ratios on both vision and language. Results indicate that a higher vision masking ratio leads to improved performance. We hypothesize that the raw signals are highly redundant for image, and a higher masking ratio is needed to facilitate representation learning. The noteworthy difference from previous work (Zhao et al. 2023a) is that we achieve better performance at a higher text masking ratio. Our interpretation is that with a more profound integration of vision and language, the model can more easily predict masked text tokens with the aid of vision.
The number of experts and the selection of top-$k$ are crucial aspects of MoE design, as they determine the model’s parameters, computational complexity, and performance. Figure 5 clearly demonstrates that performance deteriorates as the number of selected experts decreases from 2 to 1. When $k = 1$, increasing the number of experts can actually lead to a decrease in performance, which is more evident in retrieval tasks. When $k = 2$, increasing the number of experts leads to corresponding improvements in the performance of both VQA and retrieval tasks, with a more significant improvement observed in the retrieval task.

Pre-training Tasks We explore the use of different pre-training tasks for masked signal modeling in Table 5. Experiments reveal that MLM with a high masking ratio is sufficient for learning the interaction between vision and language. The addition of MIM further improves the results by reducing bias, as observed in (Kwon et al. 2023). Pre-training with MIM alone results in a minimal fusion between vision and language. We hypothesize that text descriptions are typically coarse-grained and may not offer significant assistance in fine-grained vision reconstruction. Simultaneously masking both modalities and performing MIM and MLM is not recommended. This task reduces the amount of vision and language information available, which in turn increases the difficulty of MLM and MIM, resulting in performance decline.

We further explore more pre-training tasks under the same pre-training GPU hours in Table 6. Pre-training on MIM and MLM achieves better results in both retrieval tasks and understanding tasks, thereby demonstrating the efficiency of Masked Signal Modeling. Performance on NLVR task is provided in Appendix.

Deep FFN We compare different designs of FFN in the deep layers in Table 7. Modality-shared FFN performs better than modality-specific MoE in the deep layers, as deep features require more alignment between modalities. Using a soft router can align different modalities while obtaining more modality-specific information, thereby further enhancing performance compared to deeper architecture. When we set $L_{aux} = 0$, there is a noticeable decline in the model’s

---

### Table 4: Comparison with state-of-the-art large-size models on VQA, NLVR2, MSCOCO and Flickr30K.

<table>
<thead>
<tr>
<th>Model</th>
<th>#Images</th>
<th>VQA test-dev</th>
<th>VQA test-std</th>
<th>NLVR2 dev</th>
<th>NLVR2 test-P</th>
<th>COCO TR@1</th>
<th>COCO IR@1</th>
<th>Flickr30K TR@1</th>
<th>Flickr30K IR@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>VinVL-Large (Zhang et al. 2021)</td>
<td>8.9M</td>
<td>76.52</td>
<td>76.60</td>
<td>82.67</td>
<td>83.98</td>
<td>75.4</td>
<td>58.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BLIP-CapFilt. (Li et al. 2022)</td>
<td>129M</td>
<td>78.25</td>
<td>78.32</td>
<td>82.15</td>
<td>82.24</td>
<td>81.2</td>
<td>64.1</td>
<td>97.2</td>
<td>87.5</td>
</tr>
<tr>
<td>BLIP-Large (Li et al. 2022)</td>
<td>129M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>82.4</td>
<td>65.1</td>
<td>97.4</td>
<td>87.6</td>
</tr>
<tr>
<td>Uni-PercieverMoE-L (Zhu et al. 2022)</td>
<td>44.1M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>74.7</td>
<td>58.3</td>
<td>94.1</td>
<td>83.7</td>
</tr>
<tr>
<td>FILIP-Large (Yao et al. 2022)</td>
<td>340M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>78.9</td>
<td>61.2</td>
<td>96.6</td>
<td>87.1</td>
</tr>
<tr>
<td>Prismer-Large (Liu et al. 2023)</td>
<td>12.7M</td>
<td>78.4</td>
<td>78.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT (Wang et al. 2022a)</td>
<td>800M</td>
<td>75.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>77.0</td>
<td>59.9</td>
<td>95.3</td>
<td>84.9</td>
</tr>
<tr>
<td>ALIGN-Large (Jia et al. 2021)</td>
<td>1.8B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SimVLM-Large (Wang et al. 2022b)</td>
<td>1.8B</td>
<td>79.32</td>
<td>79.56</td>
<td>84.13</td>
<td>84.84</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SimVLM-Huge (Wang et al. 2022b)</td>
<td>1.8B</td>
<td>80.03</td>
<td>80.34</td>
<td>84.53</td>
<td>85.15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Florence-Huge (Yuan et al. 2021)</td>
<td>900M</td>
<td>80.16</td>
<td>80.36</td>
<td>-</td>
<td>-</td>
<td>81.8</td>
<td>63.2</td>
<td>97.2</td>
<td>87.9</td>
</tr>
<tr>
<td>EVE-Large (Ours)</td>
<td>4M</td>
<td>79.25</td>
<td>79.20</td>
<td>84.03</td>
<td>84.69</td>
<td>82.5</td>
<td>65.2</td>
<td>96.3</td>
<td>86.3</td>
</tr>
<tr>
<td>EVE-Large (Ours)</td>
<td>16M</td>
<td>80.17</td>
<td>80.18</td>
<td>85.63</td>
<td>86.22</td>
<td>83.5</td>
<td>66.7</td>
<td>98.0</td>
<td>87.9</td>
</tr>
</tbody>
</table>

### Table 5: Ablation study on MIM and MLM.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>MIM</th>
<th>MLM</th>
<th>NLVR2 dev</th>
<th>NLVR2 test-P</th>
<th>Flickr30K TR</th>
<th>Flickr30K IR</th>
<th>VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>57.2</td>
<td>30.4</td>
<td>60.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>78.8</td>
<td>79.3</td>
<td>92.2</td>
<td>72.2</td>
<td>77.0</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>75.4</td>
<td>75.7</td>
<td>88.6</td>
<td>74.2</td>
<td>74.6</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>79.7</td>
<td>80.1</td>
<td>93.9</td>
<td>80.7</td>
<td>77.3</td>
</tr>
</tbody>
</table>

### Table 6: Ablation study on more pre-training tasks. All models are pre-trained with the same pre-training GPU hours.

<table>
<thead>
<tr>
<th>Pre-training Tasks</th>
<th>MIM</th>
<th>MLM</th>
<th>ITC</th>
<th>ITM</th>
<th>Flickr30K TR</th>
<th>Flickr30K IR</th>
<th>VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>94.0</td>
<td>80.9</td>
<td>76.8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>94.0</td>
<td>80.7</td>
<td>77.0</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>94.2</td>
<td>80.8</td>
<td>77.1</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>94.4</td>
<td>81.2</td>
<td>77.4</td>
</tr>
</tbody>
</table>
In this paper, we present a new multimodal foundation model EVE only pre-trained by Maksed Signal Modeling with Modality-Aware MoE which is flexible and capable of encoding different modalities in a unified manner. We accelerate pre-training speed 3.5x faster than pre-training with ITC and ITM. Additionally, it is easy to scale up with a larger model or more pre-training data. Extensive experiments demonstrate that EVE outperforms existing methods in various Vision Language downstream tasks.

**Modality Routing**

We compare the performance of the model whether use modality routing in the soft router or not in Table 8, and the results show that our proposed modality routing can help the router to distinguish the inputs of different modalities and thus achieve better performance.

**Visualization**

We use Grad-CAM (Selvaraju et al. 2017) heatmap to visualize the self-attention maps of EVE in masked signal modeling and VQA Task. Results are provided in Appendix.

**Conclusion**

In this paper, we present a new multimodal foundation model EVE only pre-trained by Maksed Signal Modeling with Modality-Aware MoE which is flexible and capable of encoding different modalities in a unified manner. We accelerate pre-training speed 3.5x faster than pre-training with ITC and ITM. Additionally, it is easy to scale up with a larger model or more pre-training data. Extensive experiments demonstrate that EVE outperforms existing methods in various Vision Language downstream tasks.

**Acknowledgements**

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61876224.

**References**


